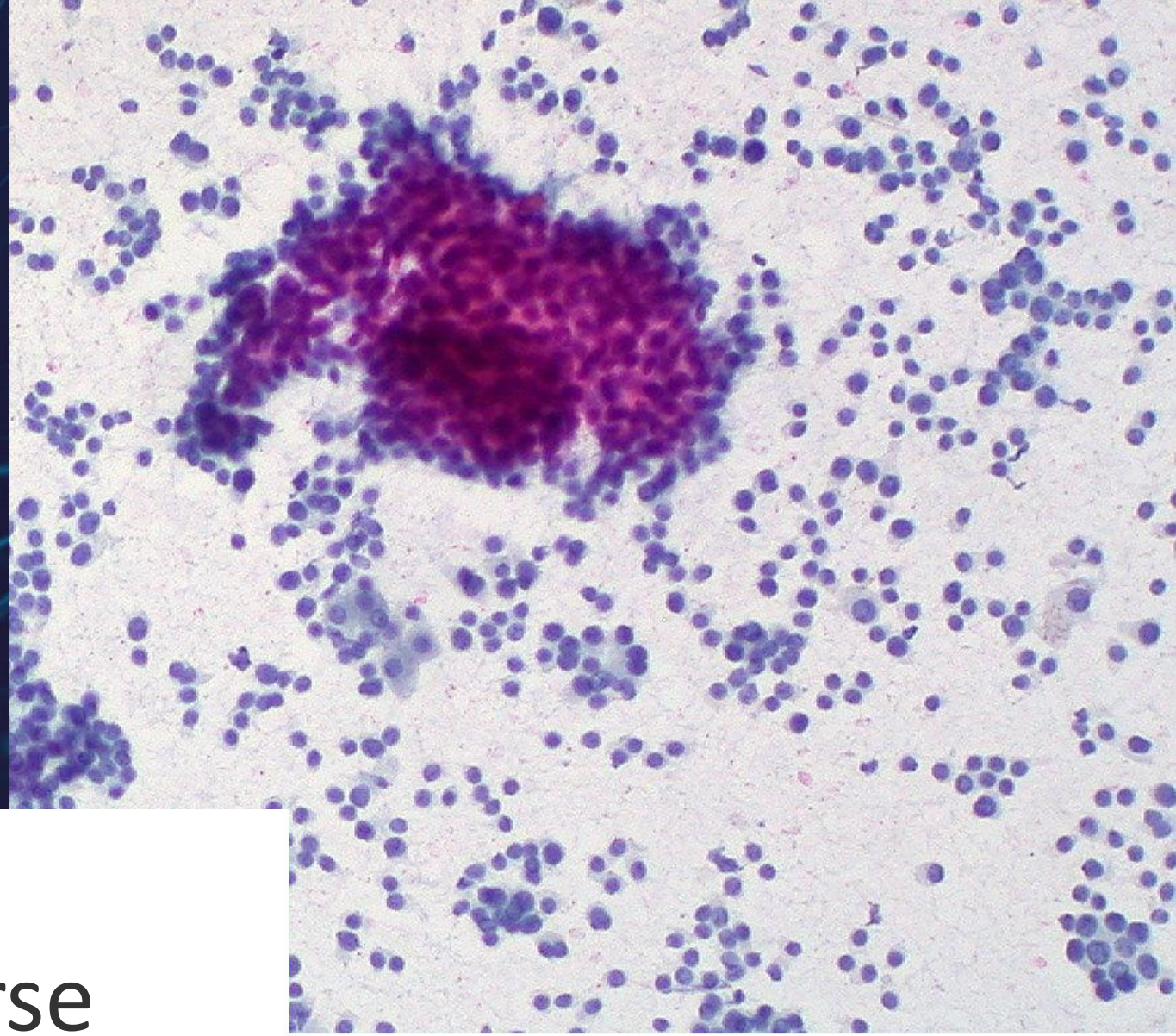


15-16 February 2021

COMETH Training course

From omics data

to tumor heterogeneity quantification

EIT Health is supported by the EIT, a body of the European Union

15 January 2021

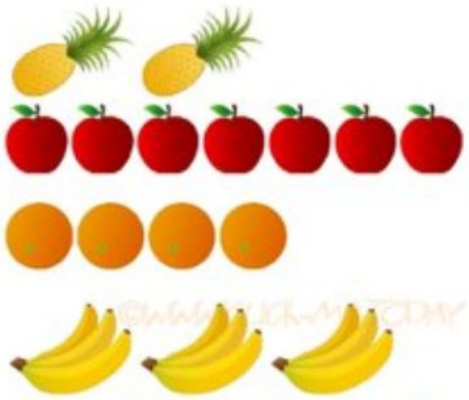
Bioinformatician point of view

Carl Herrmann and Slim Karkar

EIT Health is supported by the EIT, a body of the European Union

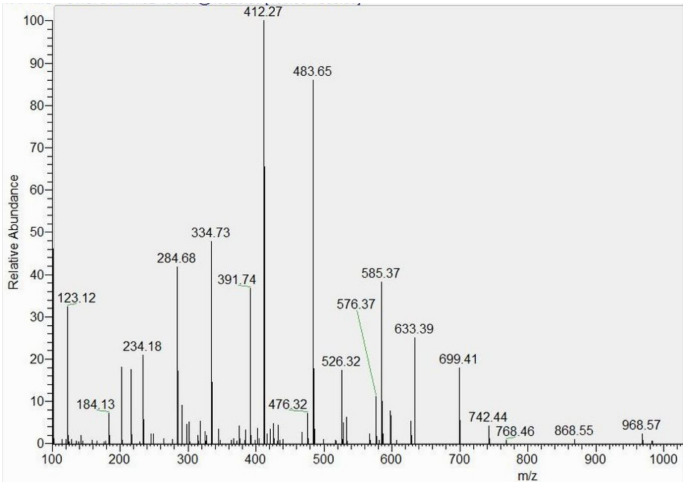
Cell-type Deconvolution

"Find the recipe!"

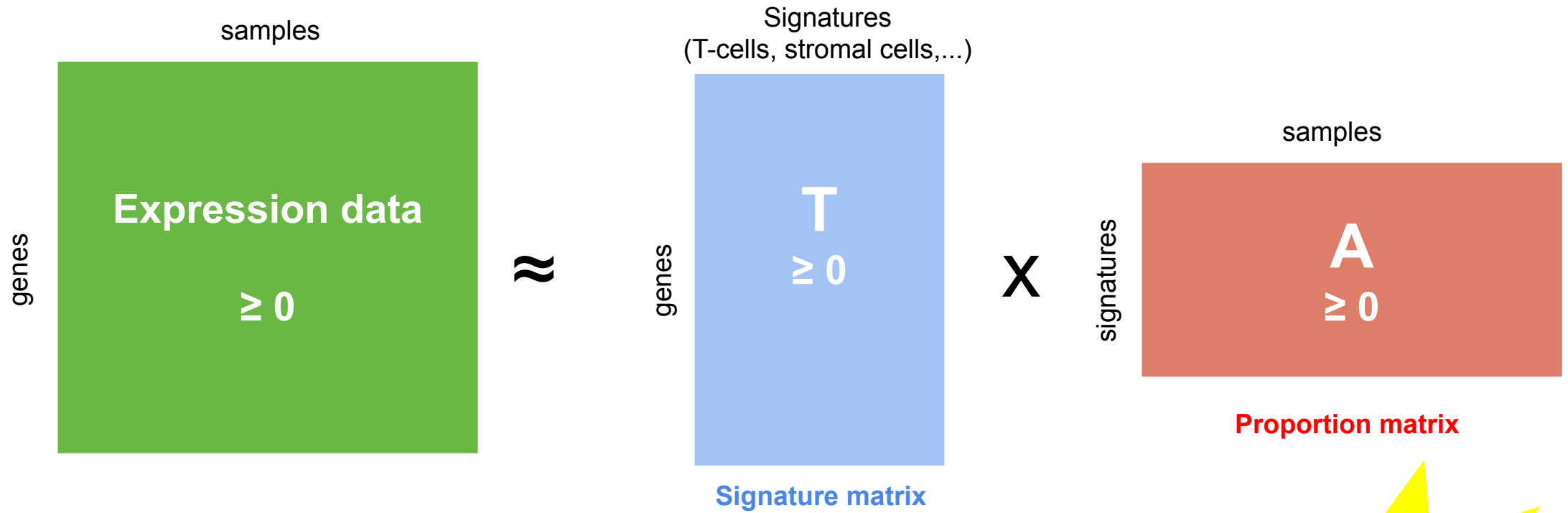


Supervised vs. unsupervised

- **Supervised methods**
  - You know the components (cherries; blueberries and bananas)
  - You know what they taste like
  - You want to know the proportions
- **Unsupervised methods**
  - You have no idea what the components are…
  - You have no idea how they taste
  - You have no idea what the proportions are
- **Semi-supervised methods**
  - You know what fruits taste like
  - But you don't know if they were used



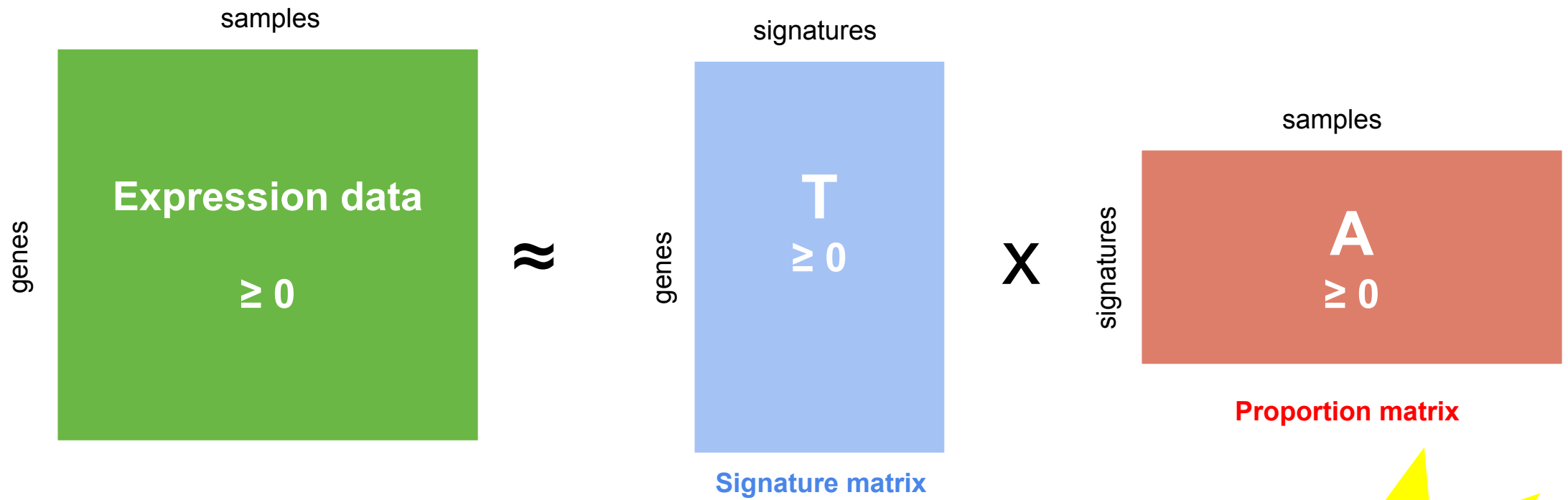
Supervised vs unsupervised methods



- **Supervised** methods:
  - Known : expression matrix ; signature matrix T
  - To be determined : proportion matrix A

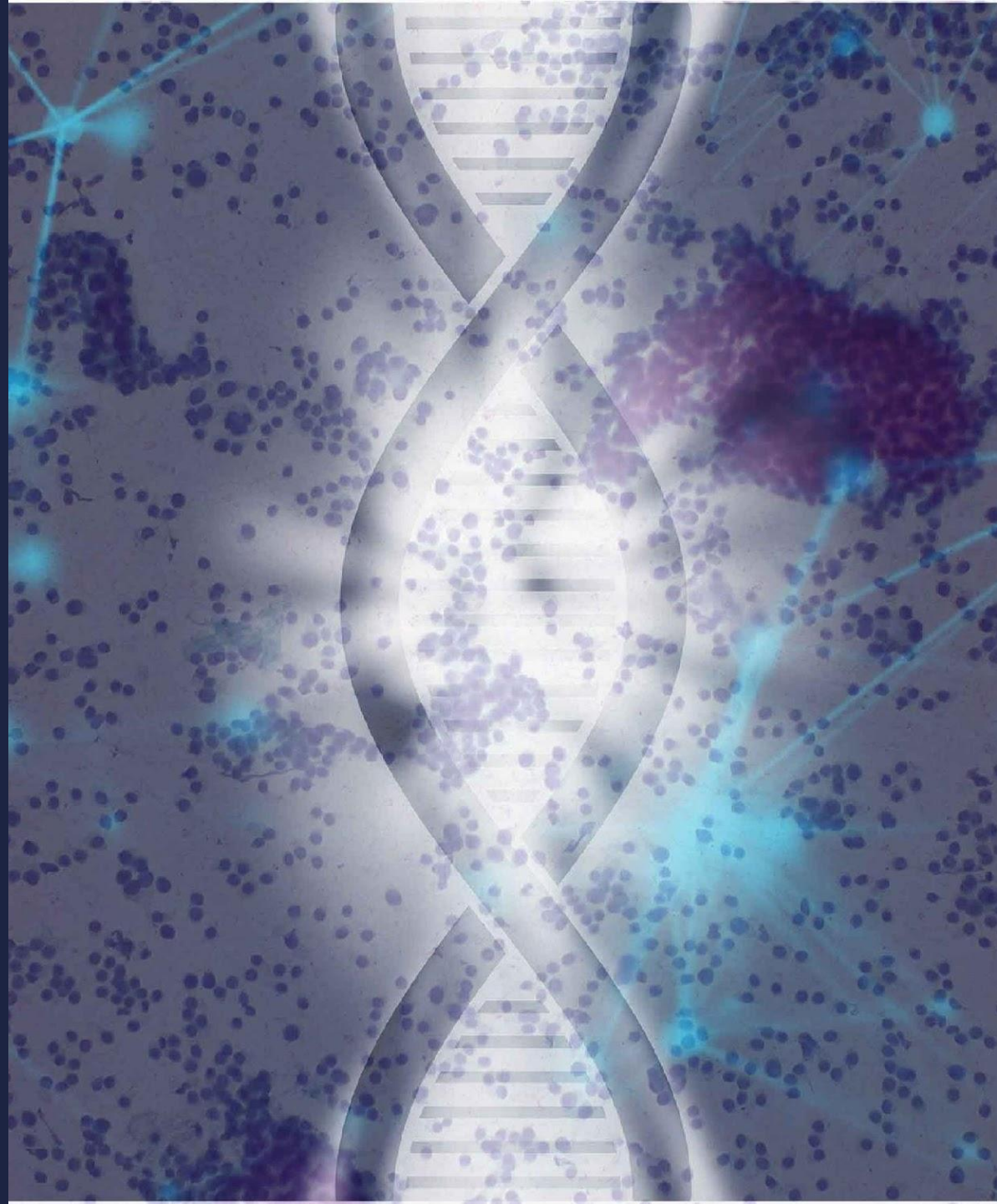
All matrices are non-negative!

Supervised vs unsupervised methods

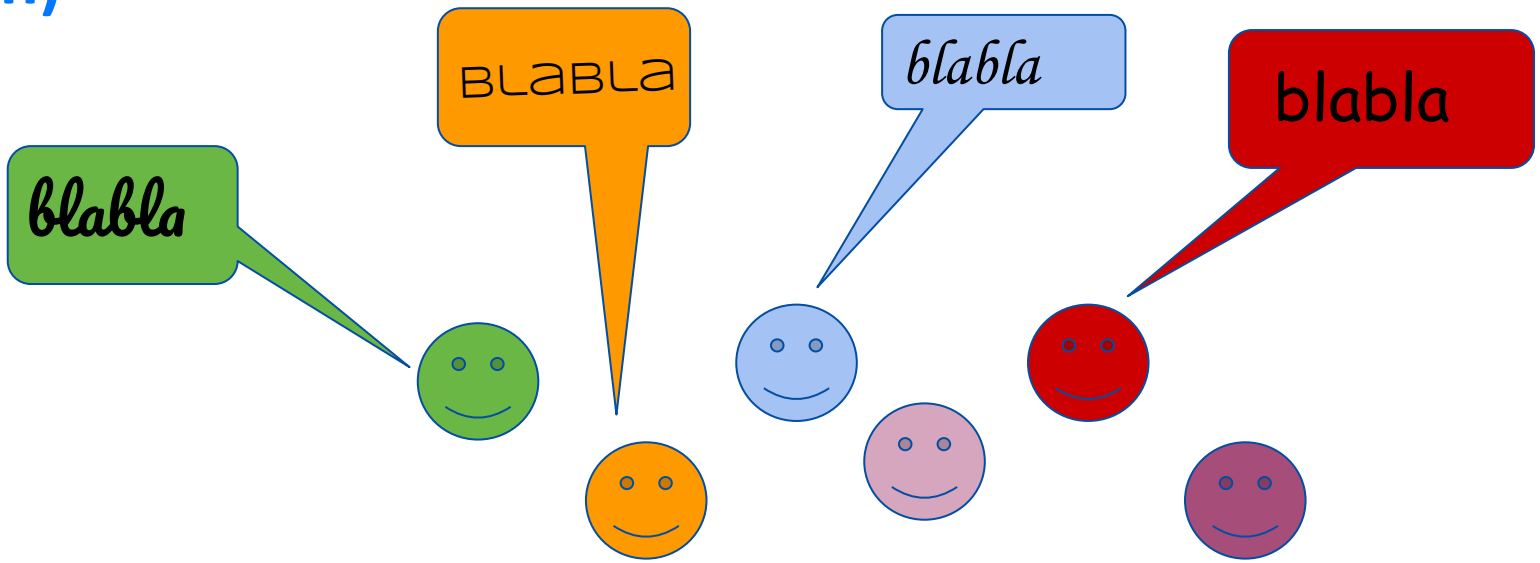


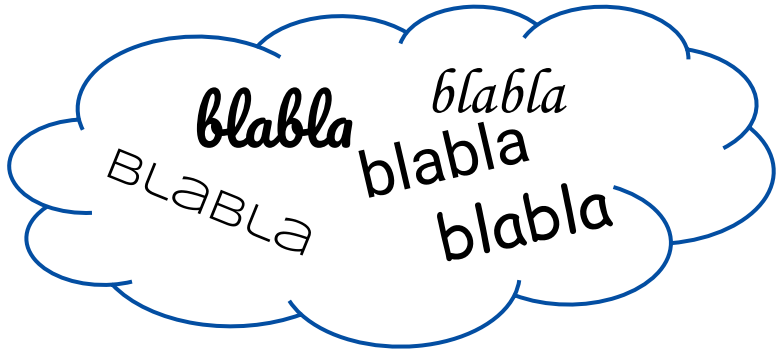
- **Unsupervised** methods:
  - Known : expression matrix
  - To be determined : T and A

Principle of supervised deconvolution



Similar problem : blind source separation problem (a.k.a. Cocktail party problem)





Supervised methods

- **Currently implemented methods**
  - CIBERSORT
  - EPIC
  - MCP-counter
  - QuantiSeq
  - TIMER
  - xCell

- They differ
  - In the **signatures** used (only immune cells / other celltypes)
  - In the **number of cell types** identified
  - In the **quantification** approach (score / proportion of all cells)
  - In the **mathematical methods** to find the A matrix (SVM-regression / least-square regression / enrichment)

Supervised methods: scores

Tool	Abbrev.	Type	Score	Comparisons	Algorithm
CIBERSORT	CBS	D	Immune cell fractions, relative to total immune cell content	Intra	ν-support vector regression
CIBERSORT abs. mode	CBA	D	Score of arbitrary units that reflects the absolute proportion of each cell type	Intra, inter	ν-support vector regression
EPIC	EPC	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression
MCP-counter	MCP	M	Arbitrary units, comparable between samples	Inter	mean of marker gene expression
quanTIseq	QTS	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression
TIMER	TMR	D	Arbitrary units, comparable between samples (not different cancer types)	Inter	linear least square regression
xCell	XCL	M	Arbitrary units, comparable between samples	Inter	ssGSEA (Hänzelmann *et al.*, 2013)

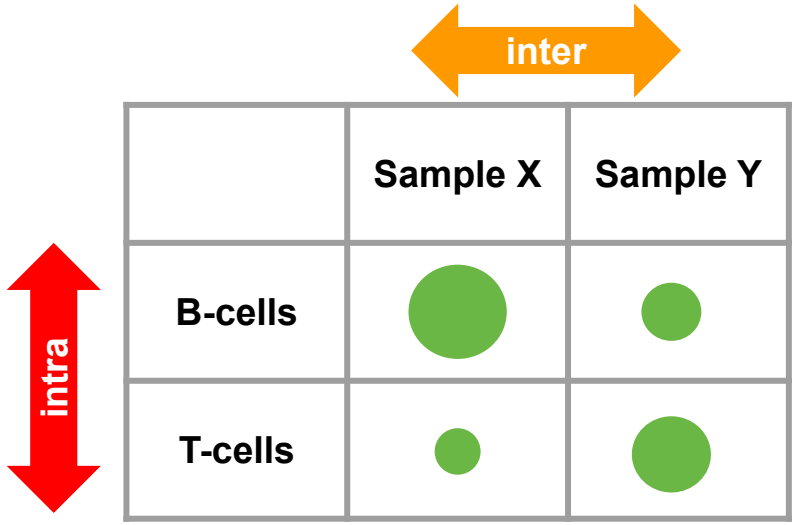
Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., … Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, *35*(14), i436–i445.

Supervised methods: scores

Tool	Abbrev.	Type	Score	Comparisons	Algorithm
CIBERSORT	CBS	D	Immune cell fractions, relative to total immune cell content	Intra	ν-support vector regression
CIBERSORT abs. mode	CBA	D	Score of arbitrary units that reflects the absolute proportion of each cell type	Intra, inter	ν-support vector regression
EPIC	EPC	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression
MCP-counter	MCP	M	Arbitrary units, comparable between samples	Inter	mean of marker gene expression
quanTIseq	QTS	D	Cell fractions, relative to all cells in sample	Intra, inter	constrained least square regression
TIMER	TMR	D	Arbitrary units, comparable between samples (not different cancer types)	Inter	linear least square regression
xCell	XCL	M	Arbitrary units, comparable between samples	Inter	ssGSEA (Hänzelmann *et al.*, 2013)

Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., … Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, *35*(14), i436–i445.

- **Intra** : you can compare relative proportions within a sample across cell types
(more B-cells than T-cells in sample X)
- **Inter**: you can compare a cell-type across samples
(more T-cells in sample X compared to sample Y)

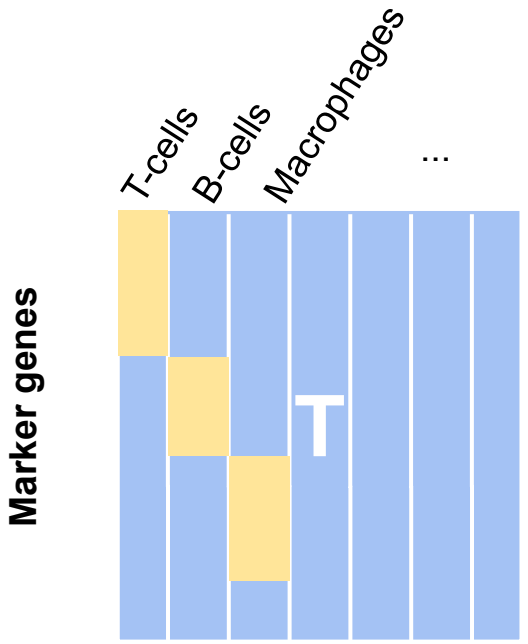


Supervised methods : signature matrix

- **Signature matrix:**
  - columns correspond to known cell types (such as immune cell types)
  - Based on identified **marker genes** which are expressed specifically for this particular cell type (using large collection of expression datasets)

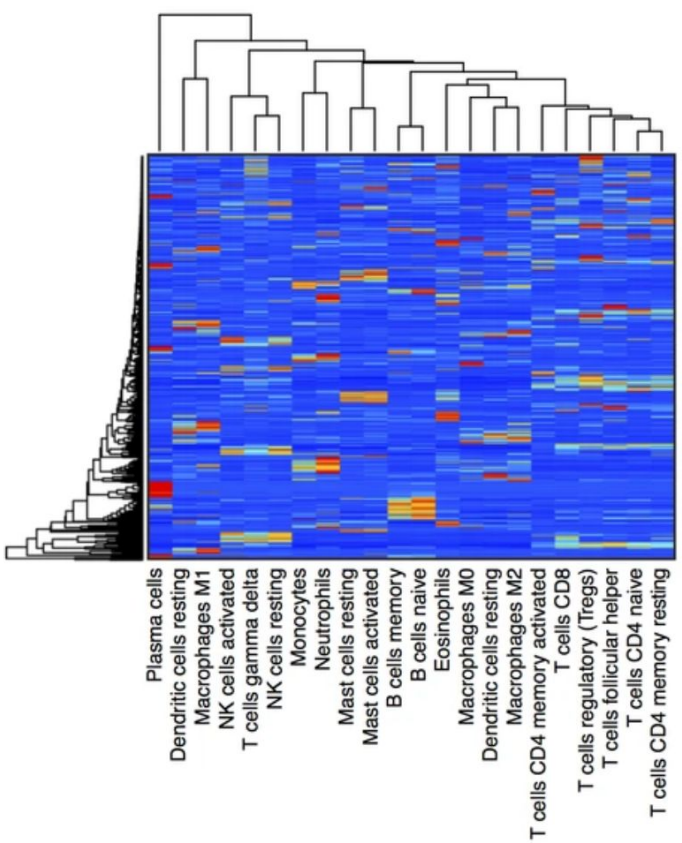
- Some methods contain a column for "other cells" allowing the identification of unknown cell types (EPIC)

- **Challenges**:
  - unspecific genes might lead to spillover of one cell types to the other (dendritic cells / B-cells)
  - Some cell types are hard to distinguish reg./non-reg. CD4+ T-cells



How to define a Signature matrix?

- **Example of the CIBERSORT method**
  - Signature matrix of 22 leukocyte populations (LM22)
  - Obtained using microarray expression data from purified cell populations
  - Differentially expressed genes between one population and all others



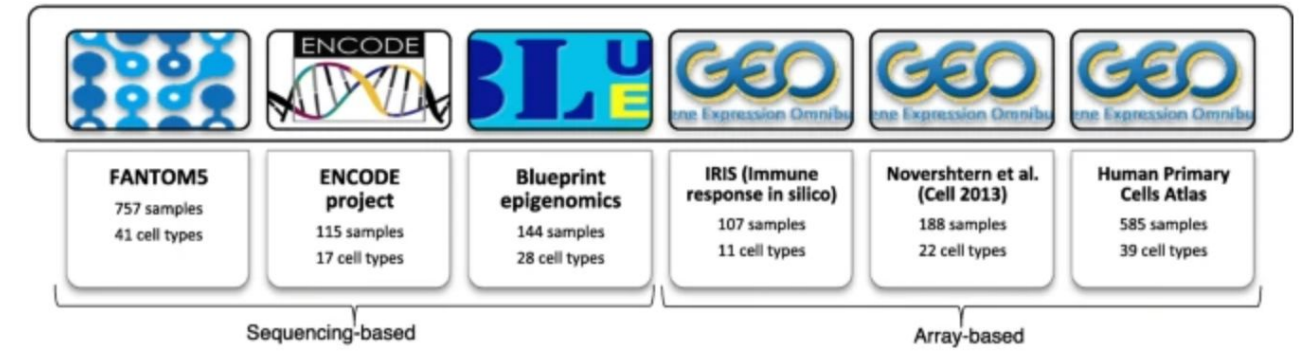
547 genes

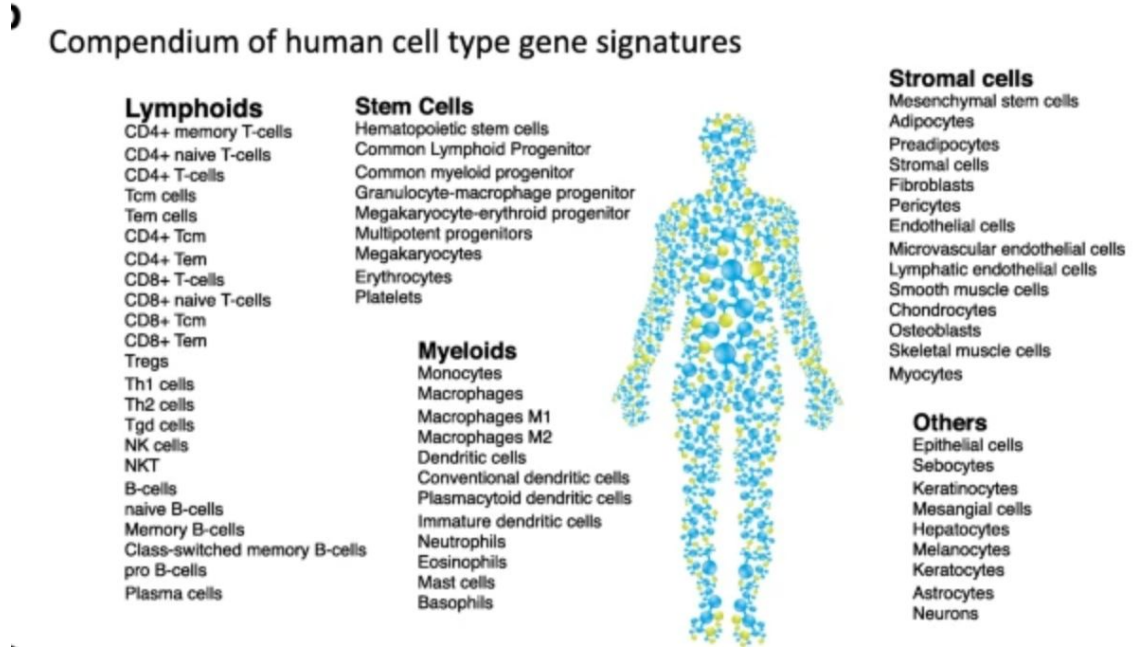
Only a subset of these genes are used for the deconvolution, depending On the datasets considered! -> feature selection

Newman, A., Liu, C., Green, M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12,** 453–457 (2015). https://doi.org/10.1038/nmeth.3337

How to define a Signature matrix?

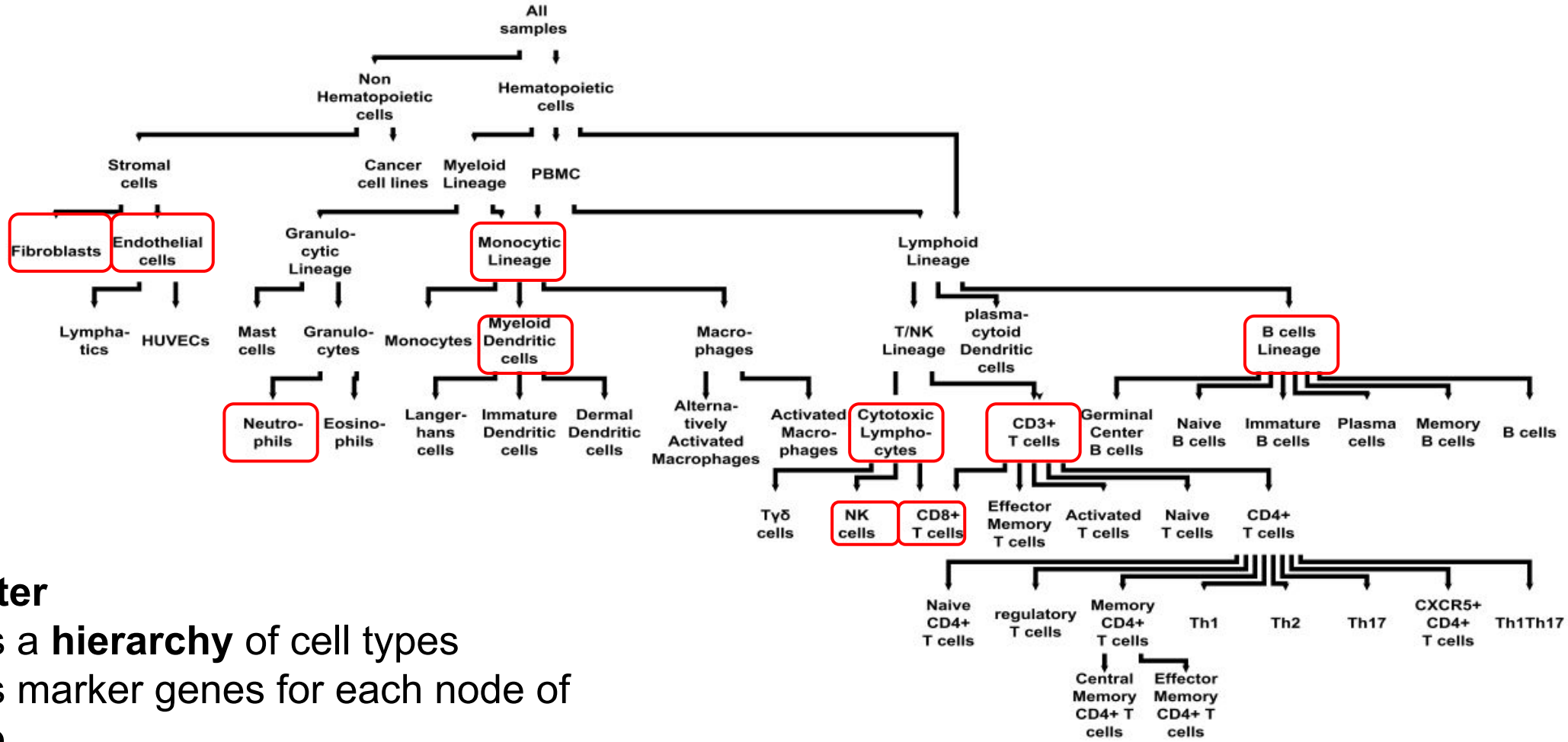
- **Example of the xCell method**
  - Large collection of 1822 expression datasets for 64 different cell types
  - Identify genes which are over-expressed in one cell type vs. all other ones
  - 6573 gene signatures for 64 cell types





Aran, D., Hu, Z., & Butte, A. J. (2017). xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, *18*(1), 1–14. https://doi.org/10.1186/s13059-017-1349-1

How to define a Signature matrix?



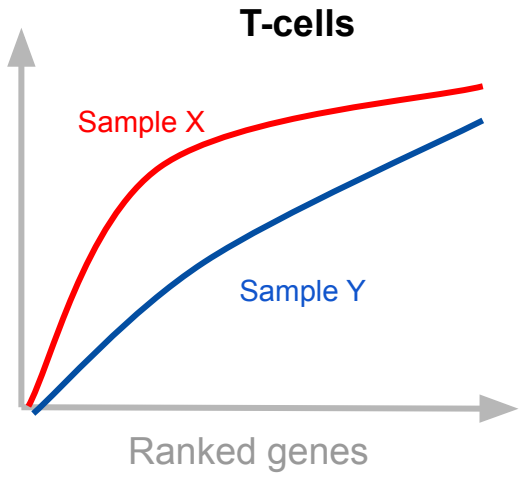
- **MCP-counter**
  - Defines a **hierarchy** of cell types
  - Defines marker genes for each node of the tree
  - 10 nodes are selected for the deconvolution process

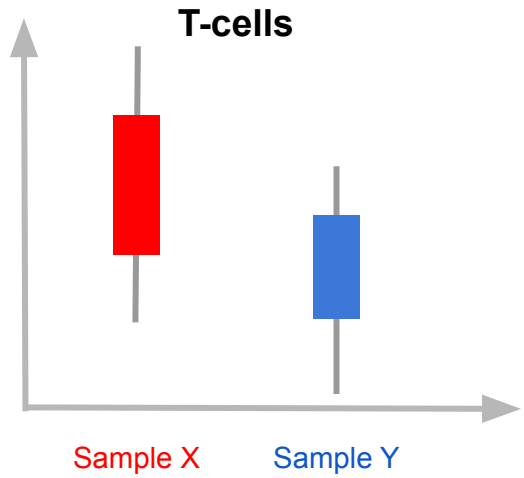
Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., … de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, *17*(1), 1–20. https://doi.org/10.1186/s13059-016-1070-5

Defining a score - performing deconvolution

Score based methods

- Compute a per sample / per cell-type score
- Use specific marker genes
- **xCell**: enrichment using ROC curve using single-cell Gene Set Enrichment Analysis (ssGSEA)
- **MCP-counter** : average log2 expression of marker genes





Defining a score - performing deconvolution

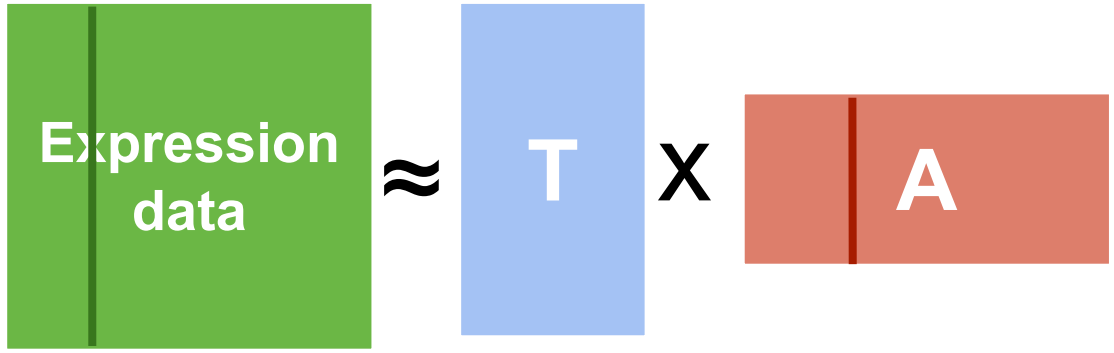
Deconvolution methods

- Gene expression in one sample = weighted sum of expression profiles in T
- Solving for the coefficients of the A matrix (= proportion matrix)

$$g_j = \sum_{i=1}^{k} t_i \cdot a_{ij}$$

Gene expression of sample *j*

Cell proportions of sample j (to be determined!)



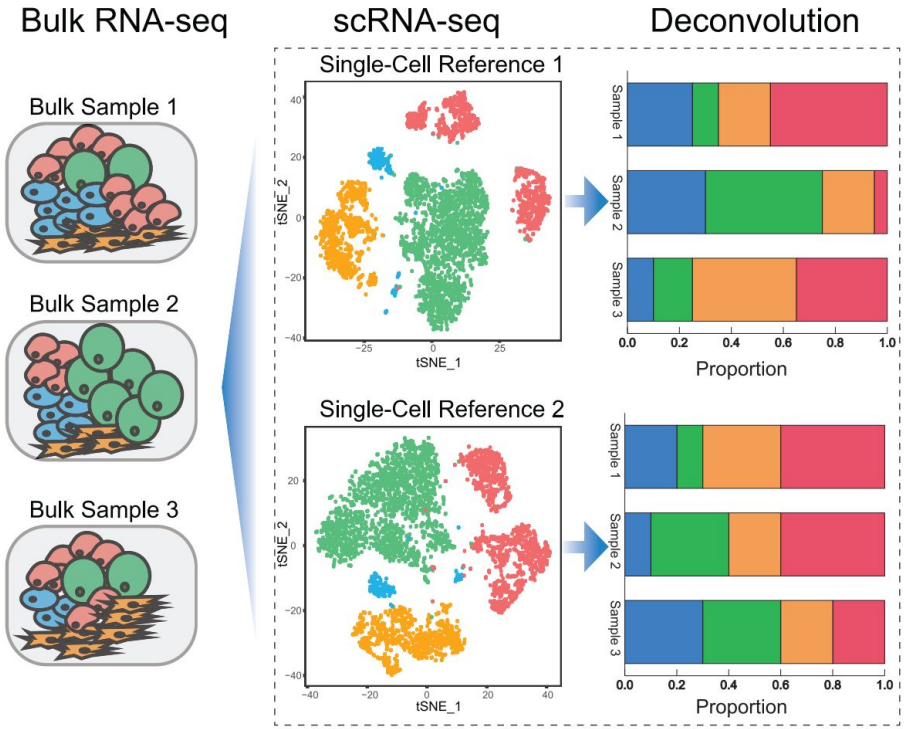
- Statistical approaches:
  - Regression : non-negative least square / contrained least-square (EPIC)
  - Support-vector machine (CIBERSORT)

Supervised methods based on single-cell reference

- Methods presented so far are based on reference expression profiles obtained from reference bulk datasets (purified cell populations)
- Increasing availability of single-cell datasets allows construction of **single-cell based references**
- **Advantage**: signal is not averaged of populations of similar cells
- **Disadvantage** : data is sparse ...

Single-cell based reference

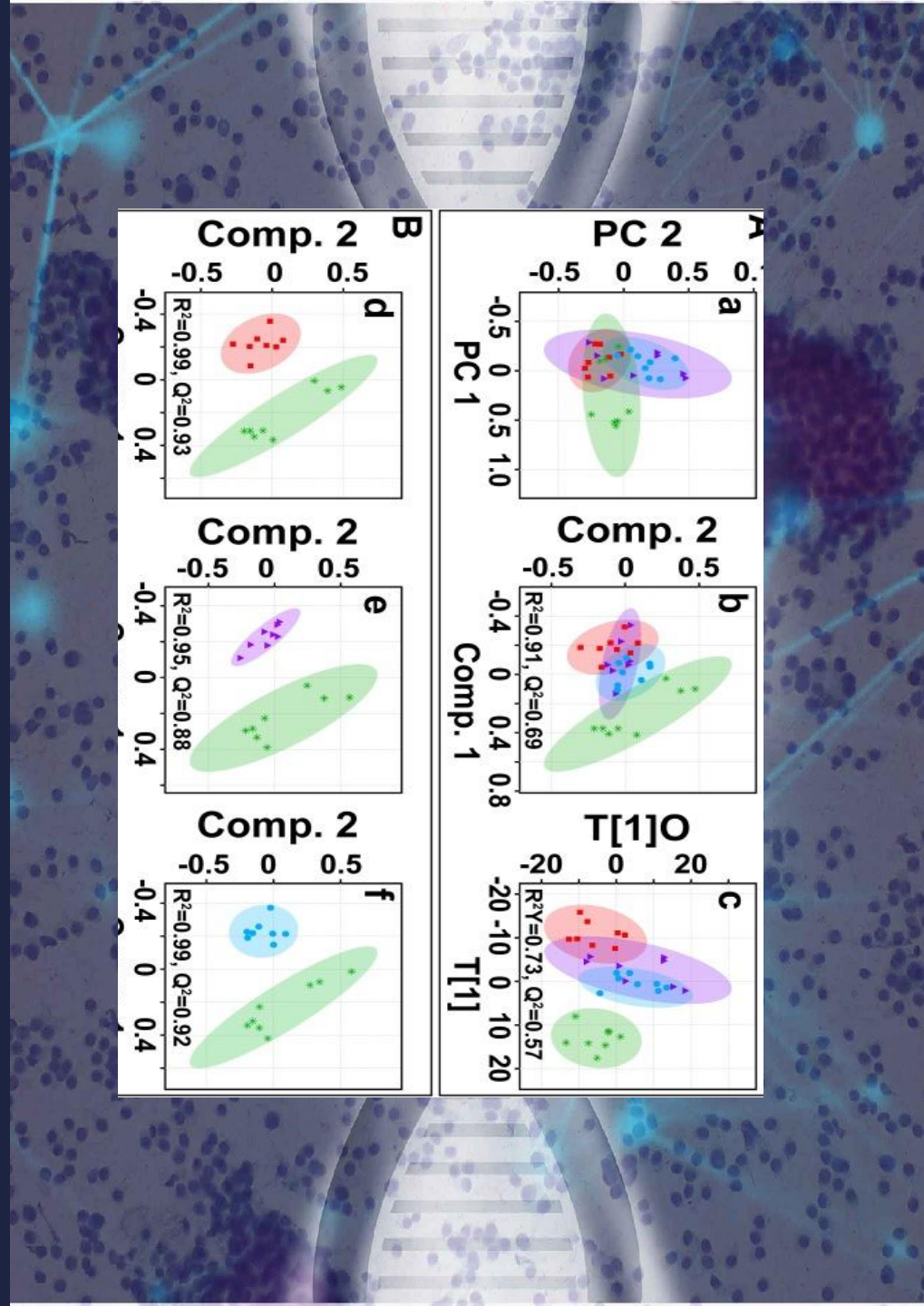
- **SCDS** : deconvolution based on **multiple single-cell references**
- Results from multiple references are aggregated and weighted



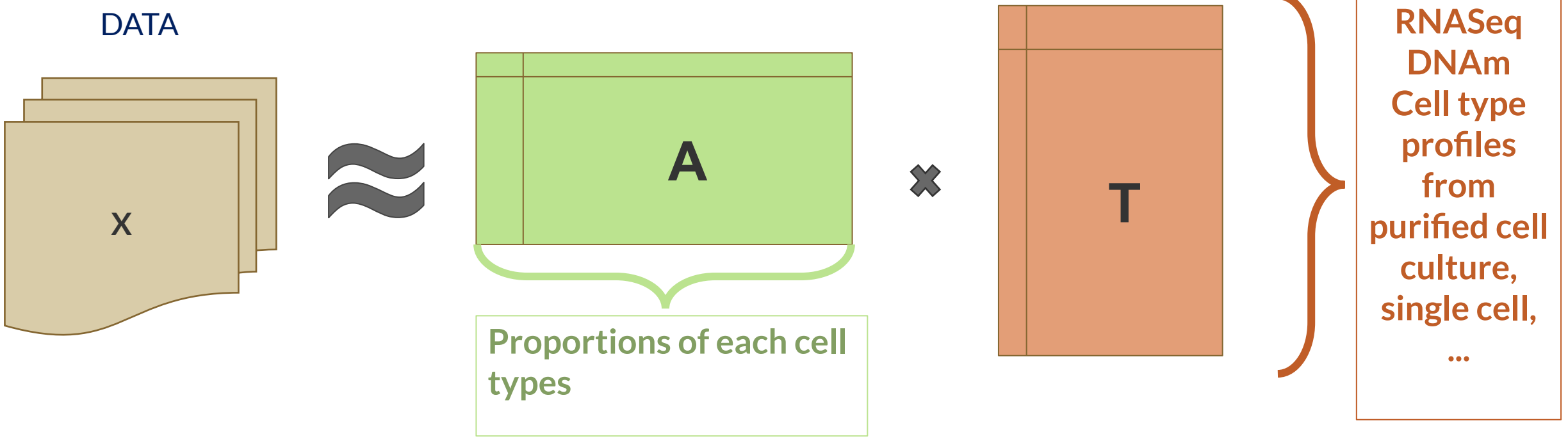
Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., & Jiang, Y. (2021). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, *22*(1), 416–427.

Unsupervised Deconvolution

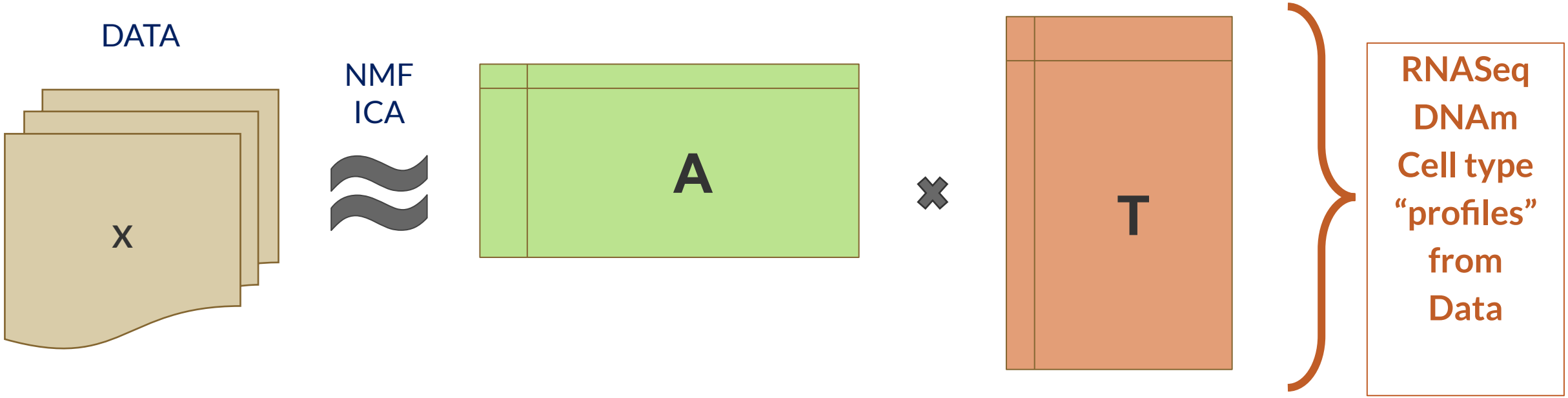
Reference-free methods



Supervised Deconvolution



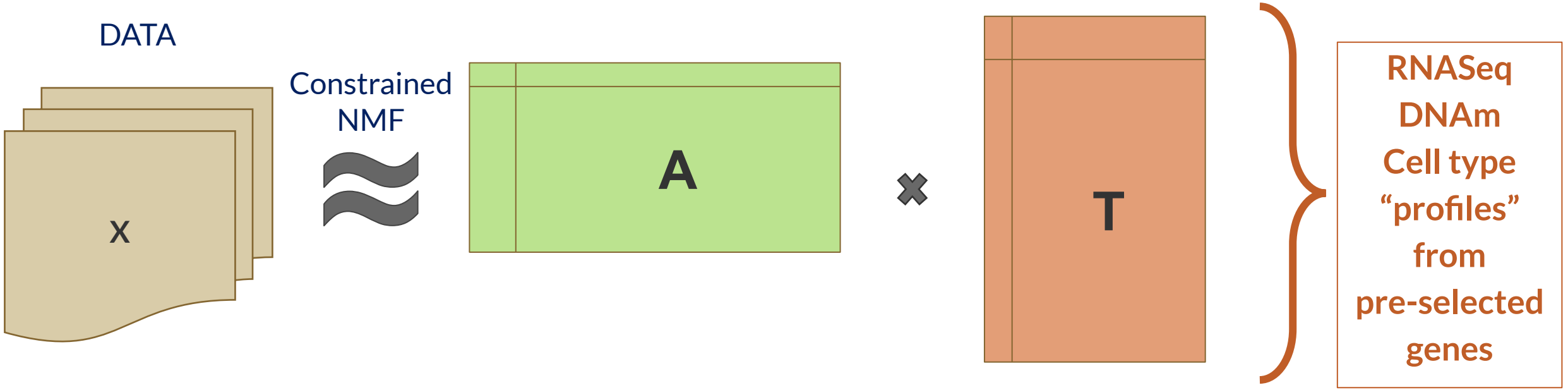
Unsupervised Deconvolution



Original space

Selected / Latent/
Embedded space

Semi-supervised Deconvolution



Updates on datasets and methods

Immune cell types quantifications on RNA seq (colorectal)

All cell types quantifications on RNA seq (breast, lung and pdac)

Reference-free
NMF (variance based feat. sel)
NMF (ICA feature selection)
ICA (no feat. sel)
ICA (ICA feature selection)

All cell types quantifications on methylome (pdac)

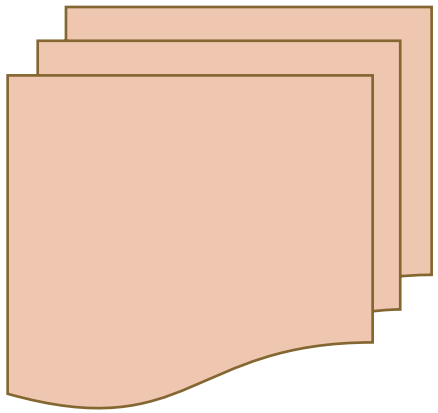
Reference-free
EDEC

All cell types quantifications on RNA seq (breast, lung and pdac)

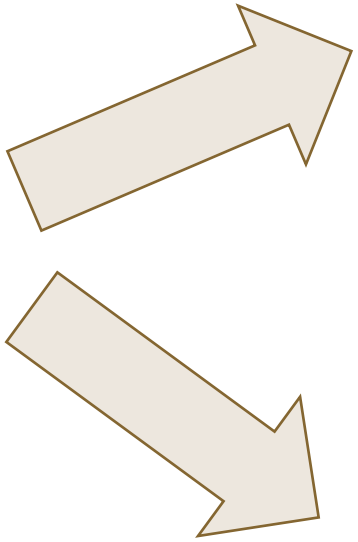
Semi-reference based (marker genes)
CellMix (NMF - KL divergence)
CellMix (NMF - euclidean distance)
CellMix (Digital sorting algorithm DSA)

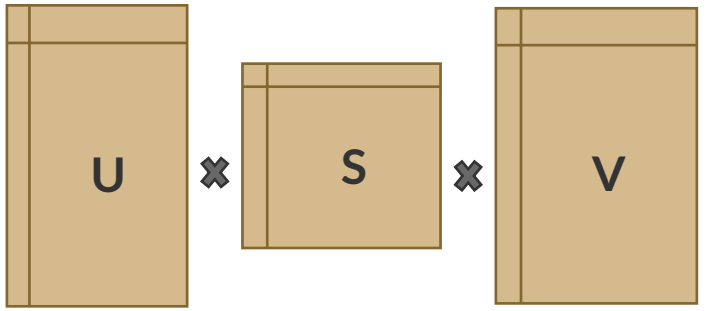
Multivariate statistics for Deconvolution

DATA

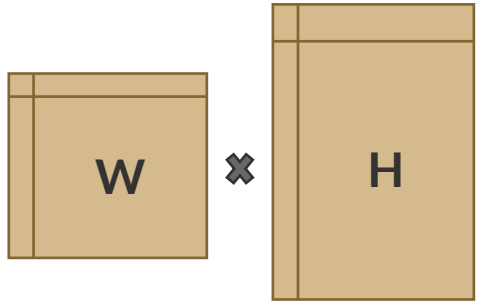


Original space





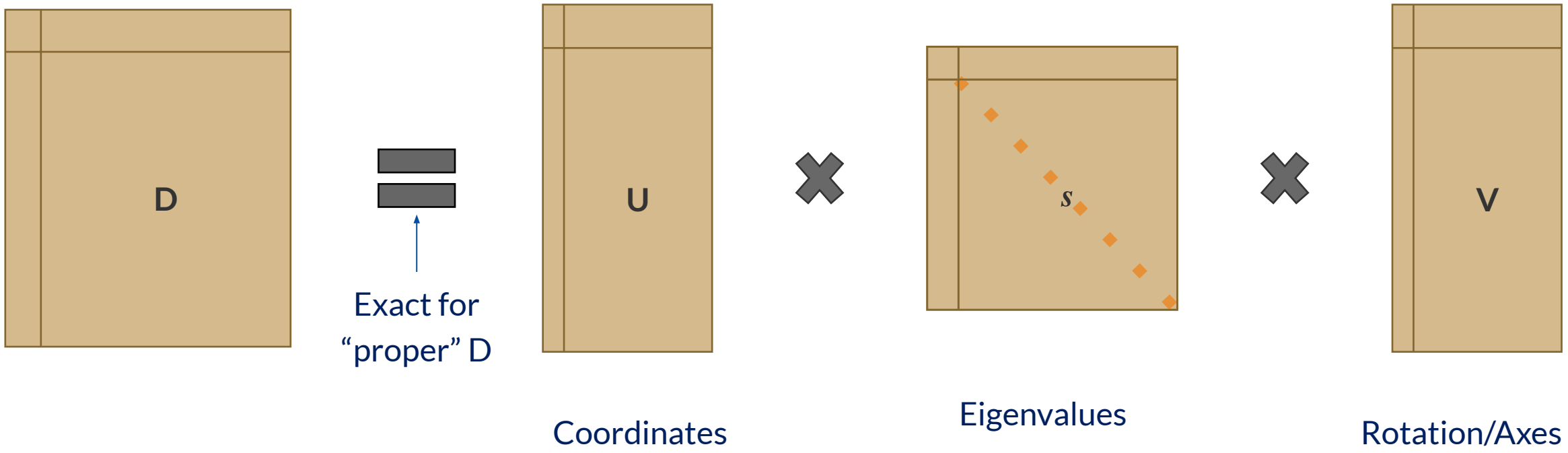
SVD



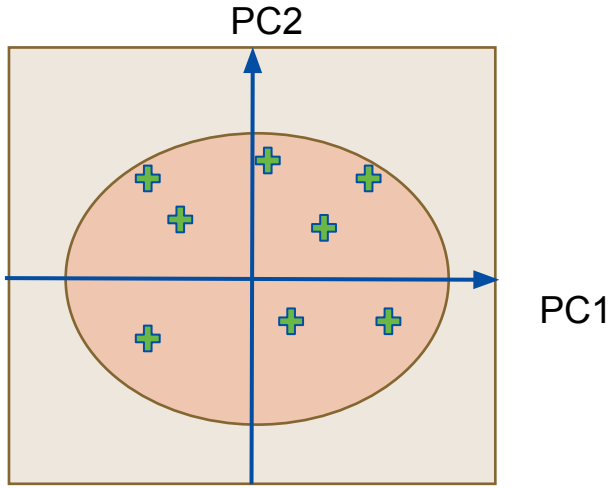
NMF
ICA

Selected / Latent/
Embedded space

Estimating number of cell types using PCA/SVD

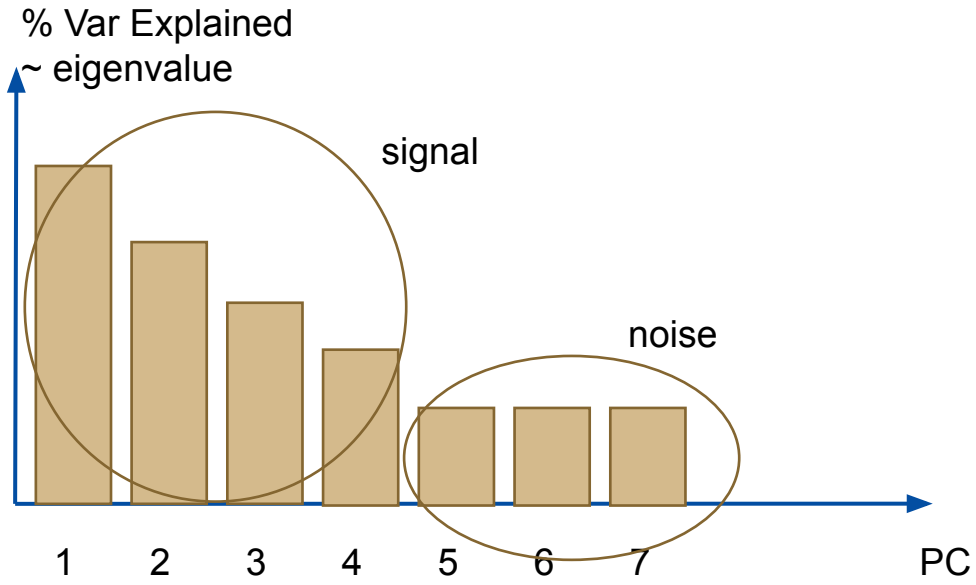


Estimating number of cell types using PCA/SVD



Scree Plot

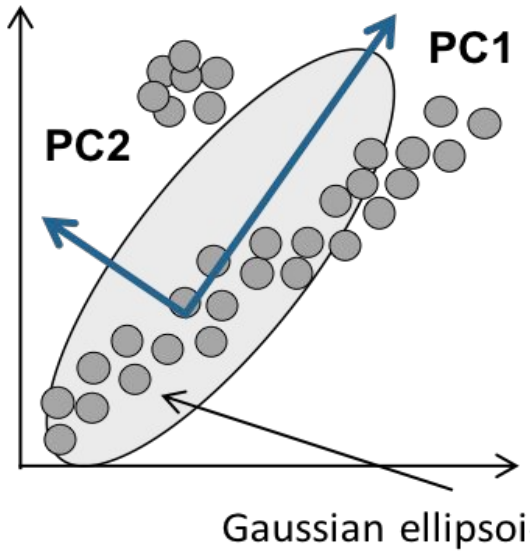
Signal : model splitting along this axes gains in var
Noise : splitting along these axes explains very little



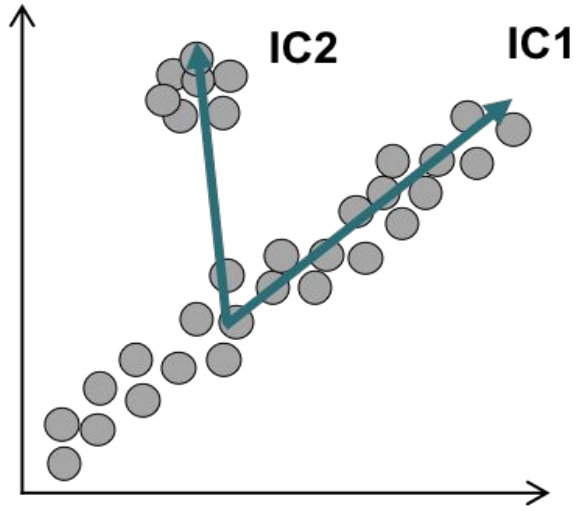
K-1 split (axis) for K clusters (“cell types” profile)
Components helps to select features (genes)

Graphical representation of dimension reduction

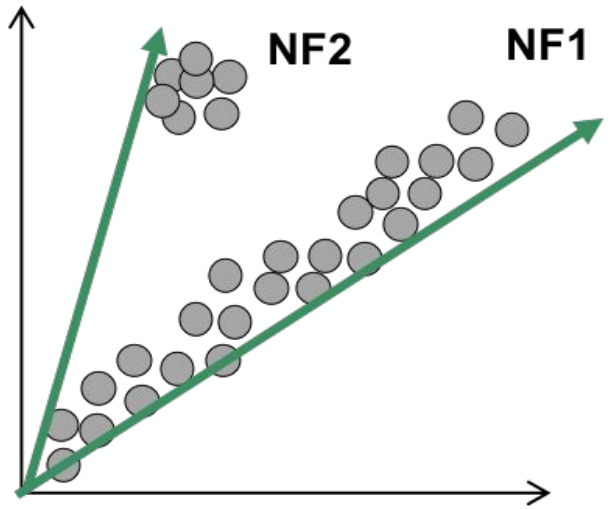
PCA does not 'sees' the data structure



Independent components are directions of **non-gaussianity**

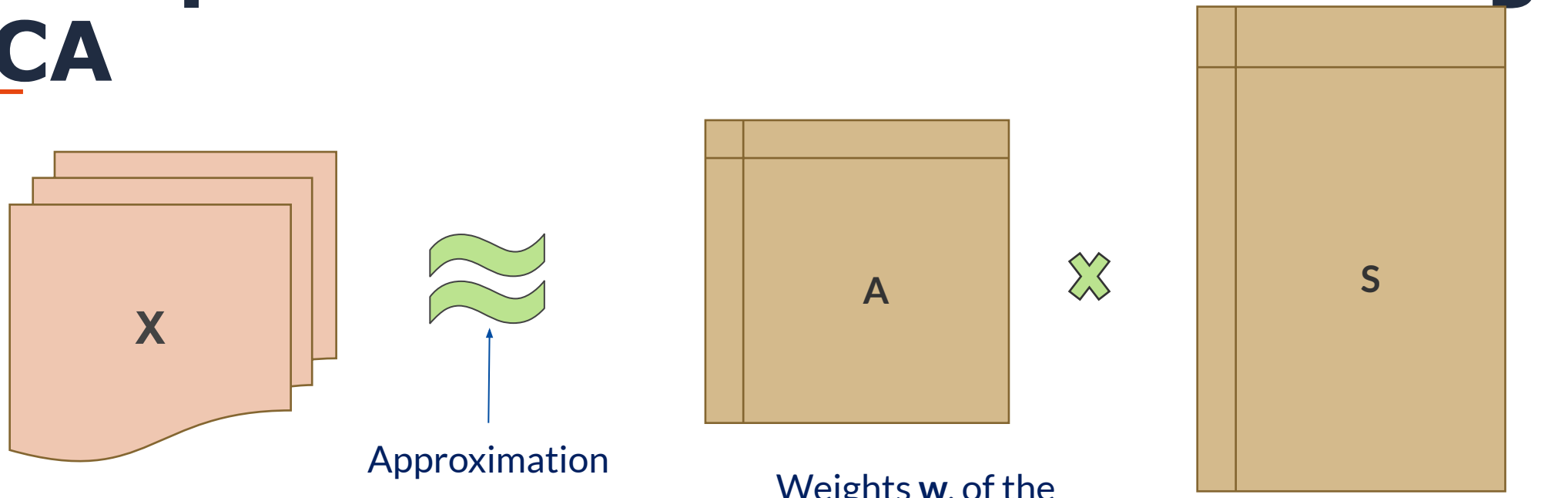


NMF components are **non-negative**



From : Introduction to DeconICA

Unsupervised Deconvolution using ICA



Data

Weights $\mathbf{w}_i$ of the components for each sample

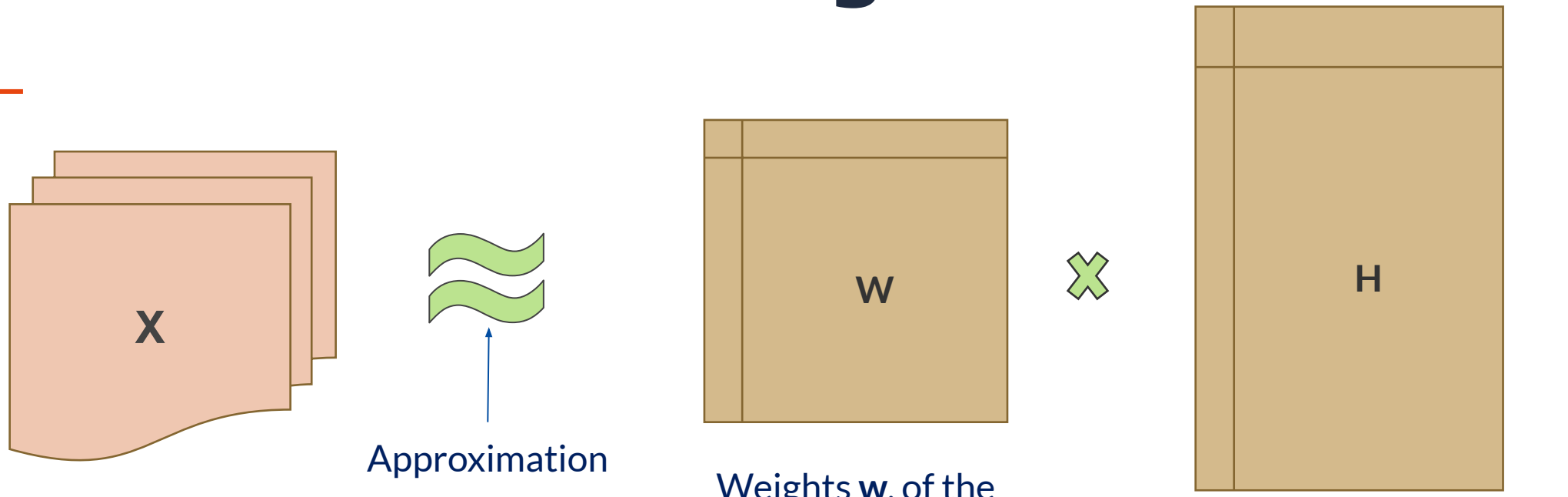
Sources : vectors of weight for genes or “metagenes”

FastICA algorithm : $A=[\mathbf{w}_1,...,\mathbf{w}_K]$; $\mathbf{S}=\mathbf{A}^T\mathbf{X}$. with ***independence*** of components $\mathbf{w}_i$

Orientation : independence can lead to **negatively oriented component**

Stabilité : $\mathbf{w}_i$ are initialized randomly ;

Deconvolution using NMF

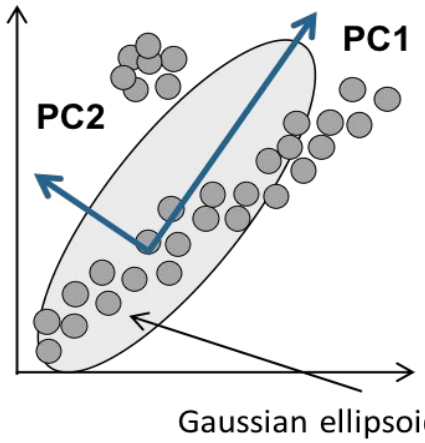


NMF : determine **W**,**H** that minimizes $f(\mathbf{W},\mathbf{H})=\frac{1}{2} \; // \mathbf{X}-\mathbf{W}^T\mathbf{H} //_F^2$
Where $// . //_F$ is Frobenius norm, Kullback-Leibler divergence - CellMix : euclidean distance

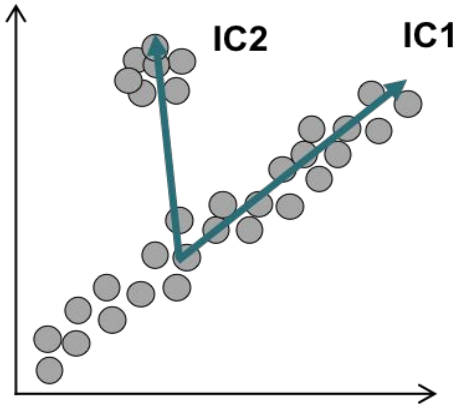
Semi-supervised : constrained on **H** to use (only) ***marker genes***

To go further into details

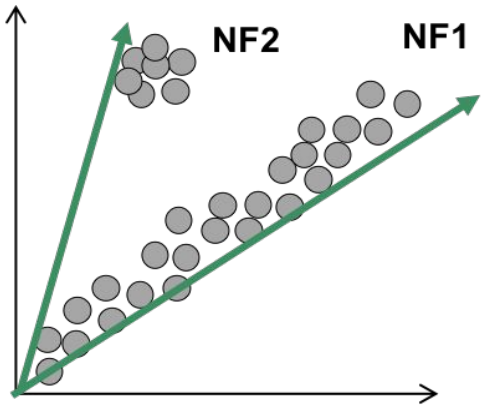
PCA does not 'sees' the data structure



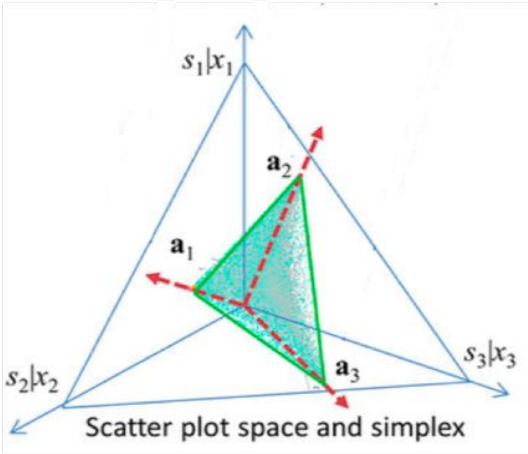
Independent components are directions of **non-gaussianity**



NMF components are **non-negative**



Fitting data to a **convex hull** shape



Graphical representation of dimension reduction & BSS methods.

PCA, ICA, NMF inspired by figures of Andrei Zinovyev, Convex hull: CC BY (Wang et al. 2016)

https://urszulaczerwinska.github.io/DeconICA/DeconICA_introduction.html

https://urszulaczerwinska.github.io/UCzPhDThesis/

Data Challenge for deconvolution

DATA CHALLENGE

Benchmark : Deconvolution from Expression and Methylation Data

https://www.codabench.org/competitions/237/?secret_key=b164d1c1-07ca-4d0c-b55f-99e68af3a343

How to participate?

(1) **Register** to the challenge on Codalab
(2) **Find** your teammates on discord
(3) **Download** the starting kit and the public datasets

CHALLENGE BEGINS
(1) **Work** in group to build prediction models
(2) **Submit** your code or results on the Codabench platform
(3) **Improve** your score
CHALLENGE ENDS

Feedback on your work to the other teams
3 slides per team - online presentation or PDF format
-Approach (1 slide)
-Results (1 slide)
-Discussion, pros & cons (1 slide)

Yuna Blum, Ligue contre le Cancer

Jérôme Cros, APHP

Clémentine Decamps, Uni Grenoble Alpes

Carl Herrmann, Medical Faculty Heidelberg

Slim Karkar, Uni Grenoble Alpes

Yasmina Kermezli, Uni Grenoble Alpes

Magali Richard, Uni Grenoble Alpes

Ashwini Sharma, Uni Grenoble Alpes

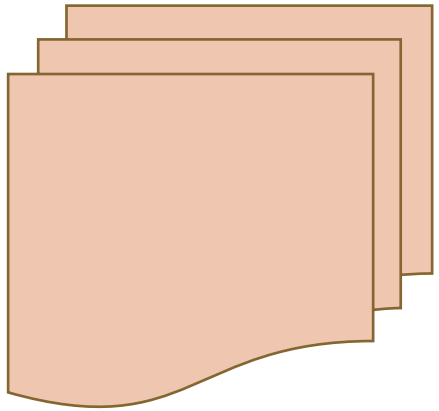
https://cancer-heterogeneity.github.io/cometh_training.html

www.eithealth.eu | info@eithealth.eu

EIT Health is supported by the EIT, a body of the European Union

Cell types profiling

scRNAseq
pure type DNAm

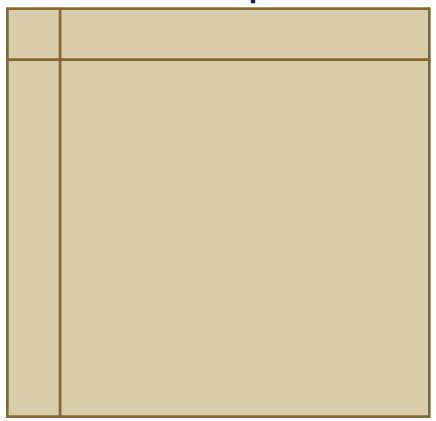


Original space

Dimension Reduction

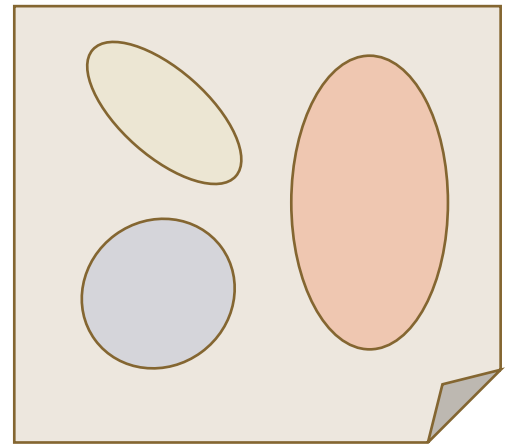
PCA
t-SNE

n Selected Features
n I/P Component



Selected
Embedded
space

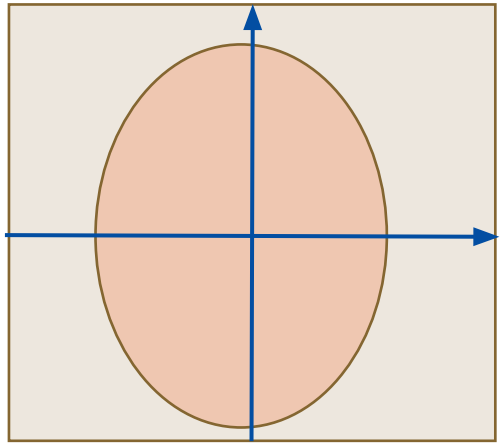
plot and/or clustering

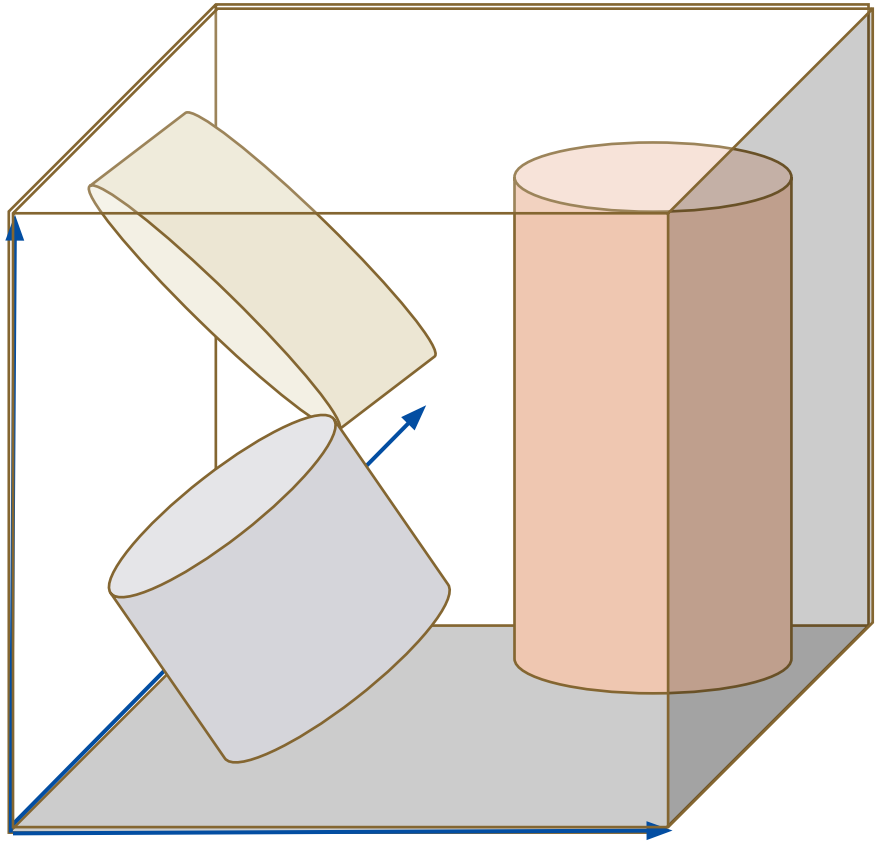


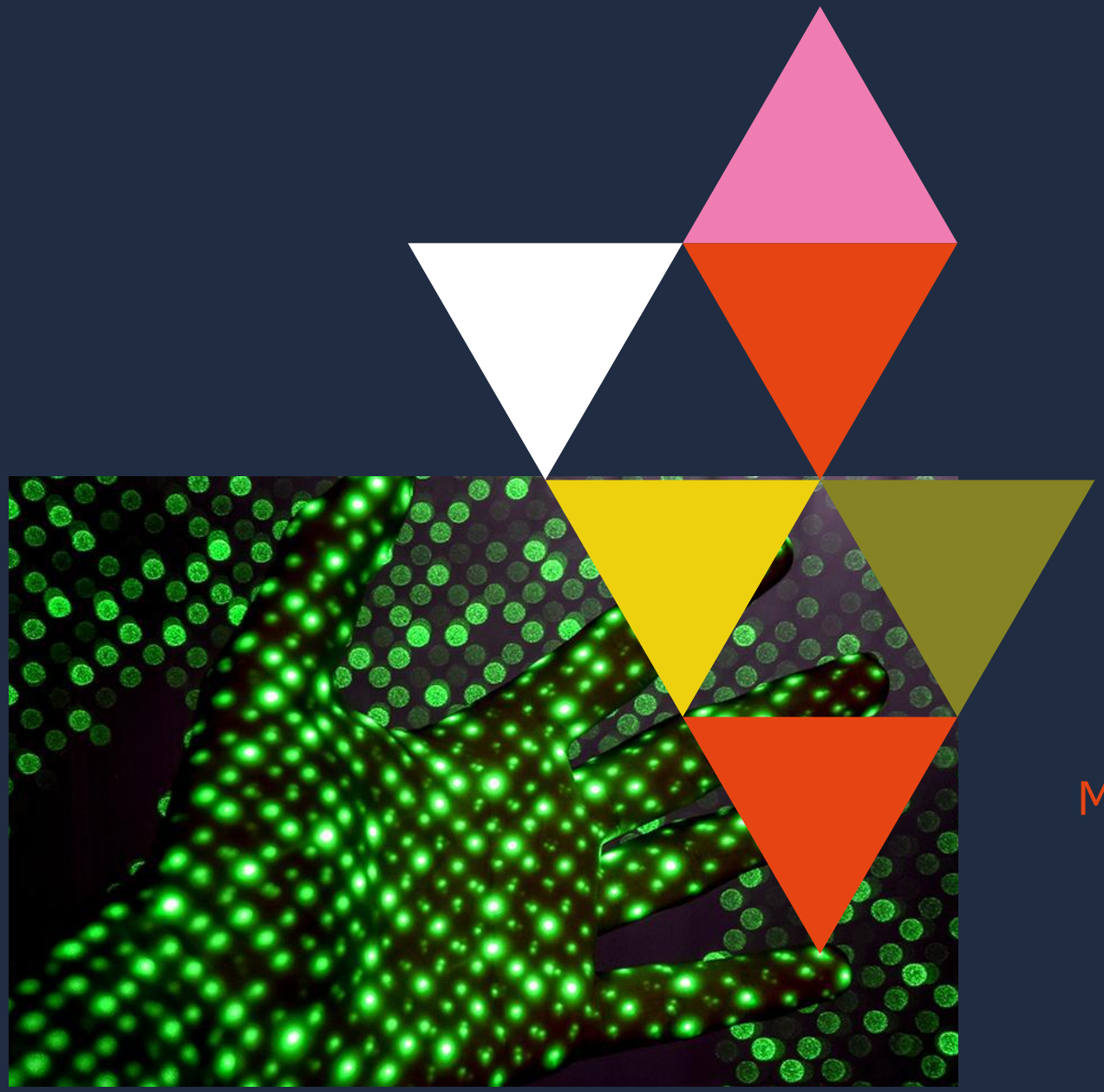
Selected
Embedded
space

Immune cell types quantifications on RNA seq (colorectal)	**Reference based (expression table)** **EPIC** **Cibersort** **Quantiseq**
All cell types quantifications on methylome (pdac)	**Reference based (expression table)** **EpiDISH**

Deconvolution using PCA/SVD



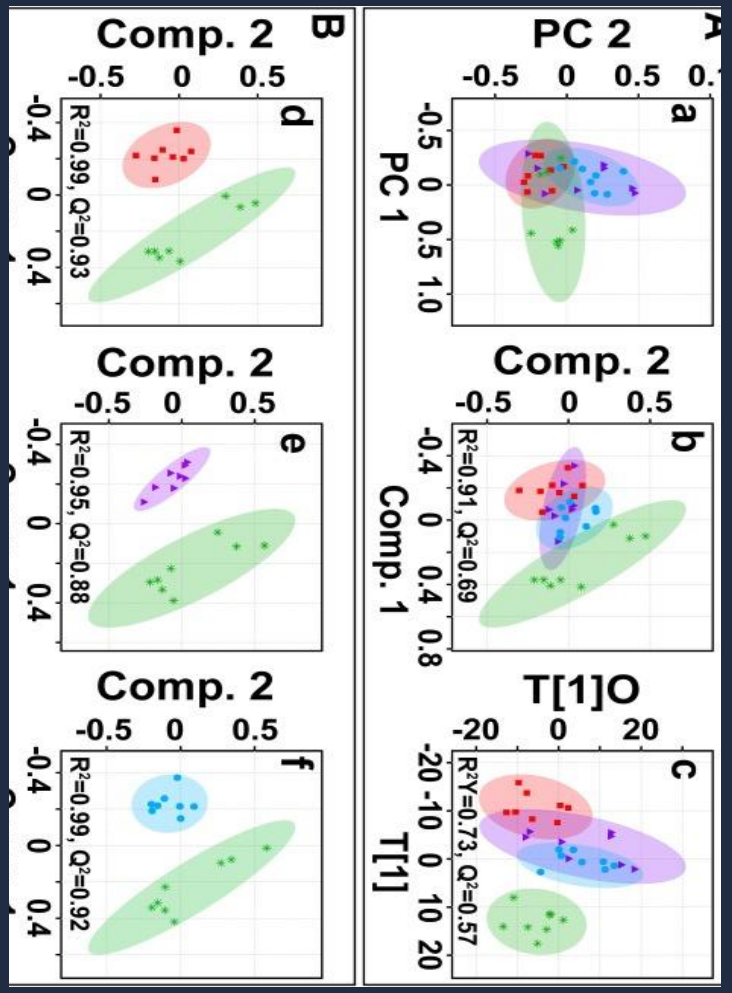




Université Grenoble Alpes

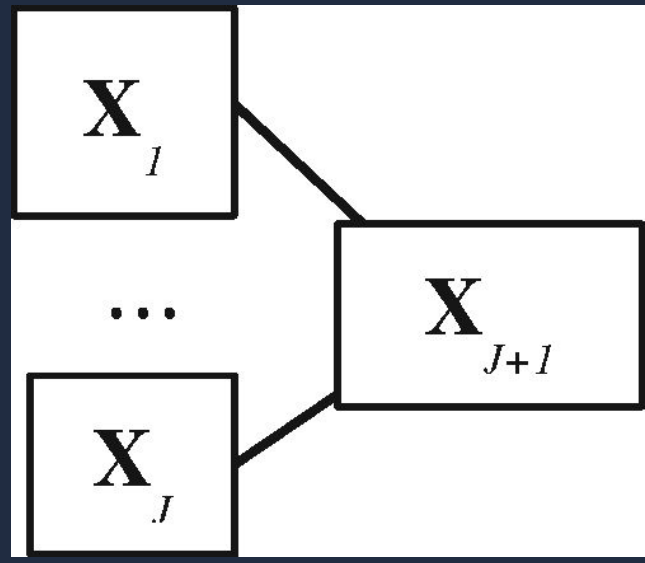
IA4Health School Practical Session

Machine Learning algorithms for prediction of cancer outcomes from multiomic data

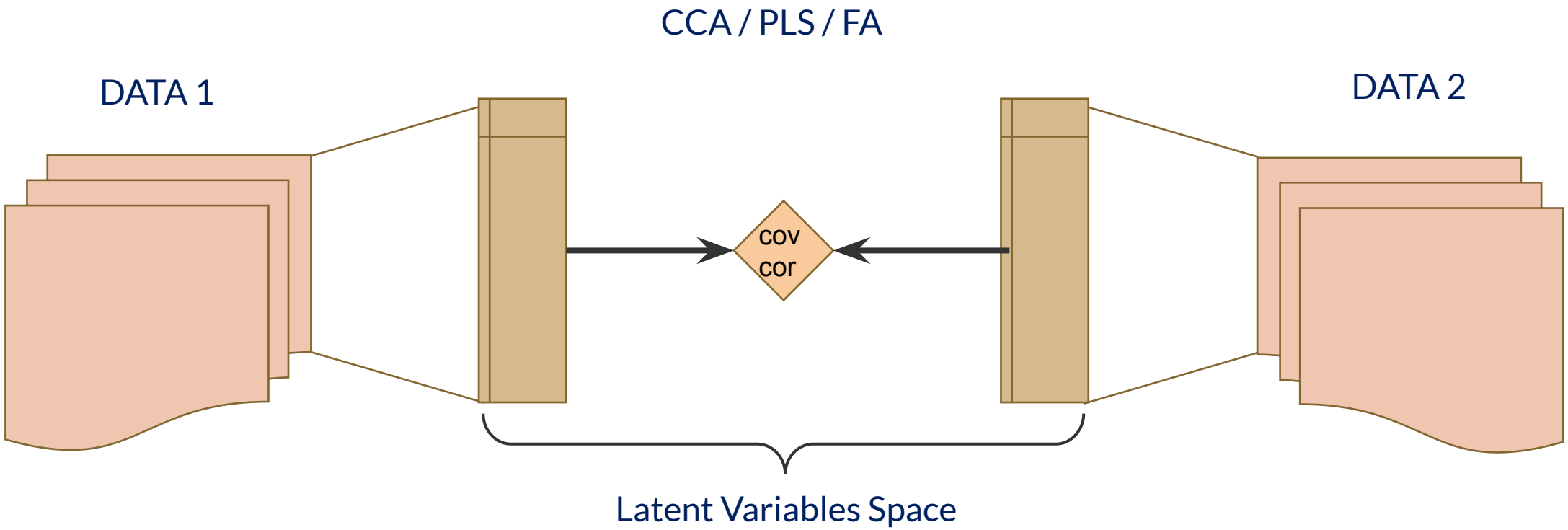


Multivariate Analysis

What are multivariate statistics ?

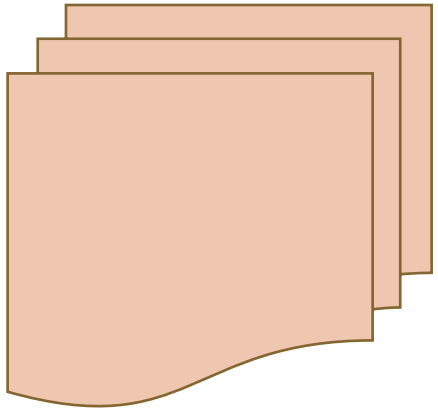


Prediction using Multivariate Statistics



Prediction using Multivariate Statistics

DATA

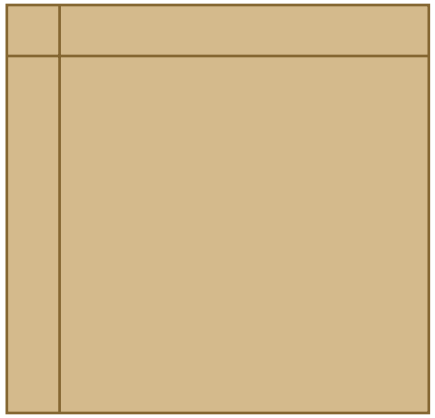


Original space

Dimension Reduction

PCA / ICA
SVD / NMF
CCA / PLS / FA

Selected Features
Latent Variables

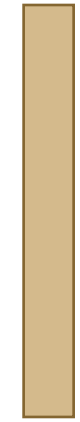


Selected / Latent/
Embedded space

Regression
ML

lm/glm/glmnet
boost/randomForest
SVM/NaiveBayes/LDA

Prediction



Prediction using Multivariate Statistics

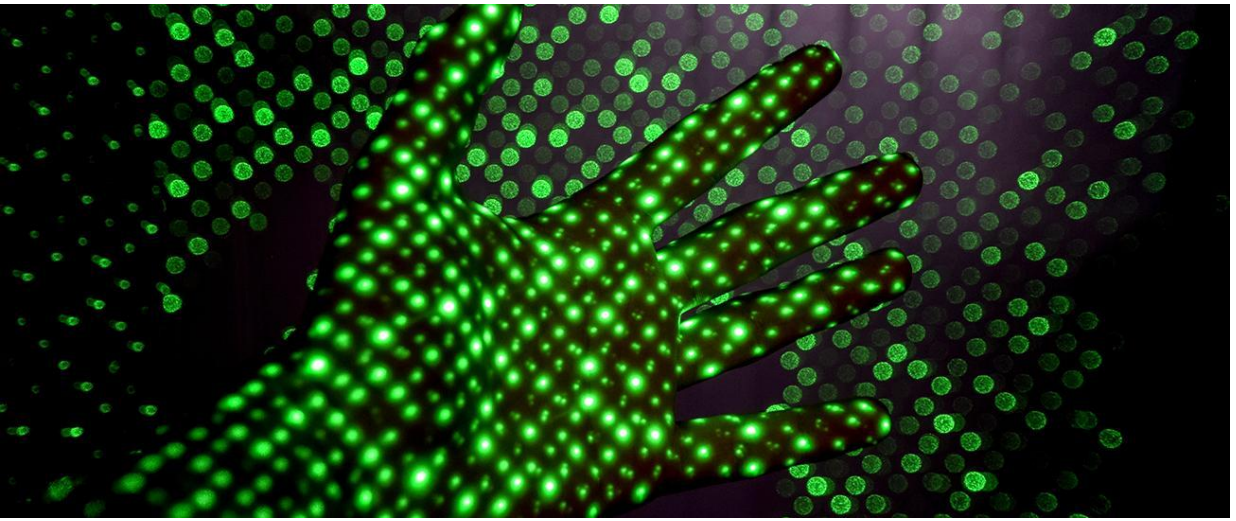
Good start :

Feature Selection + PCA

18.6 "High-Dimensional Regression: Supervised Principal Components" p.694
ESLII.pdf
More Details :
MultiOmicsDR_Table.pdf

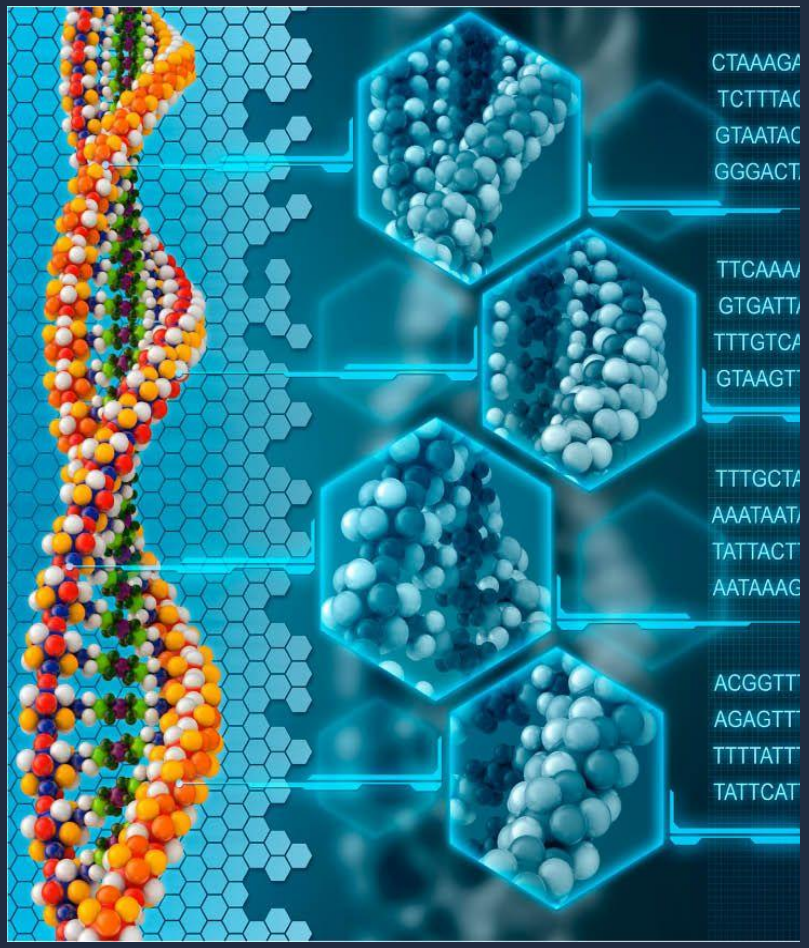
tPCA (tensors)	***Tensorial extension of PCA***	***R***
PARAFAC (tensors)	***Tensorial extension of PCA***	***R***
tensor CCA	CCA	MATLAB
sCCA	***CCA***	***R***
MCCA	CCA	NO
CCA-RLS	CCA	NO
RGCCA	***CCA***	***R***
DIABLO	***CCA***	***R***
jointNMF	NMF	MATLAB
MultiNMF	NMF	NO
EquiNMF	NMF	NO
IntNMF	***NMF***	***R***
iCell	NMF-based	MATLAB
Scikit-fusion	Matrix	python
Higher-order GSVD (HOGSVD	***SVD (Matrix tri-factorization)***	***R***
iCluster	***Gaaussian latent variable model***	***R***
funcSFA	Gaaussian latent variable model	python
JIVE	***PCA***	***R***
AJIVE	PCA	MATLAB
MCIA	***Co-Inertia***	***R***
MOFA	***Factor Analysis (FA)***	***R***
Group	***Factor Analysis (FA)***	***R***
MSFA	***Factor Analysis (FA)***	***R***

Sommaire



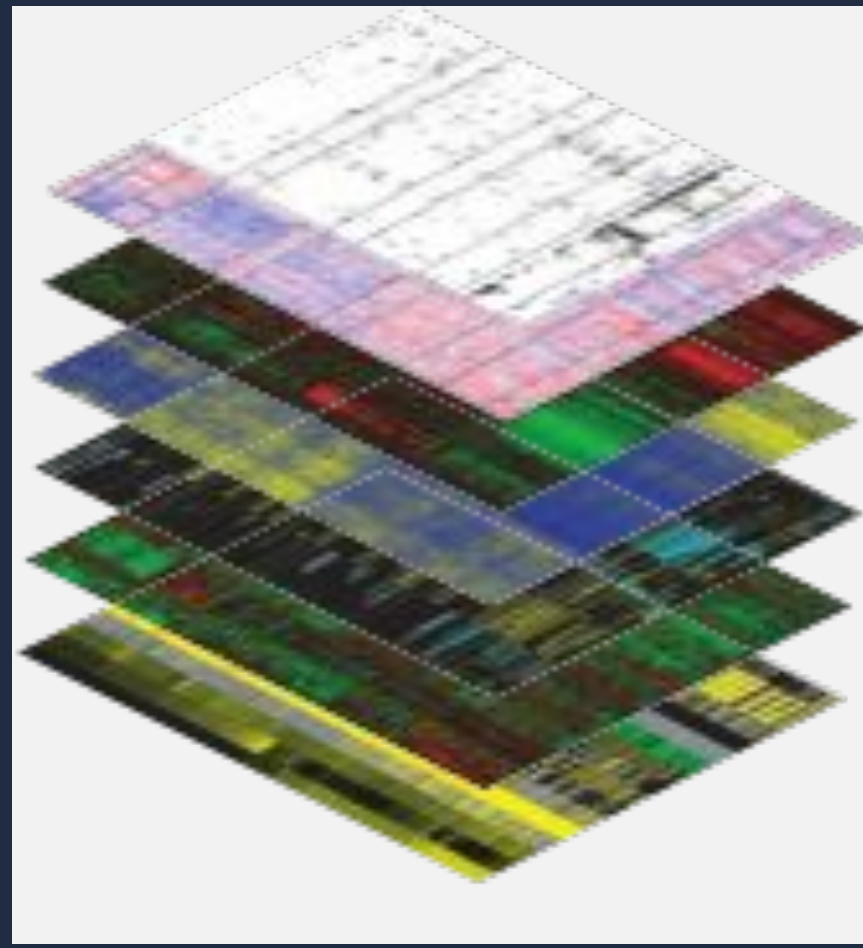
Data Visualization for Genomics Data
Heat Map
PCA
MFA
Regressions
Logistic regression
Penalized Regression
Multivariate Prediction
CCA/PLS
Regularized CCA
Sparse CCA

TP 1
Prediction of Histological classes from Expression Data
TP2
Survival prediction form Expression and Methylation Data



Visualizations of Genomic Data

Why do we need visualization ?

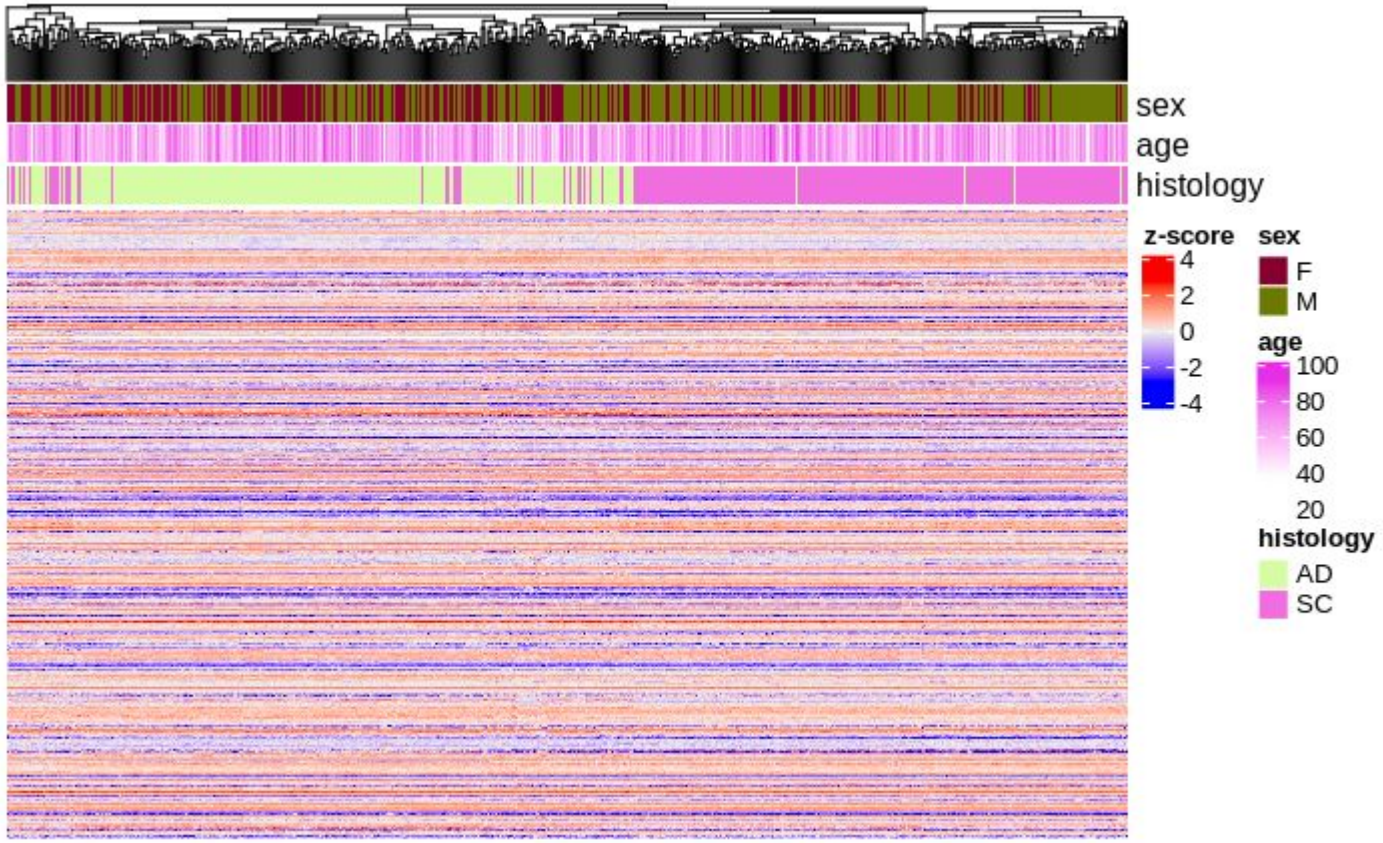


Data Visualization for Genomics Data : Heat Map and Z-score

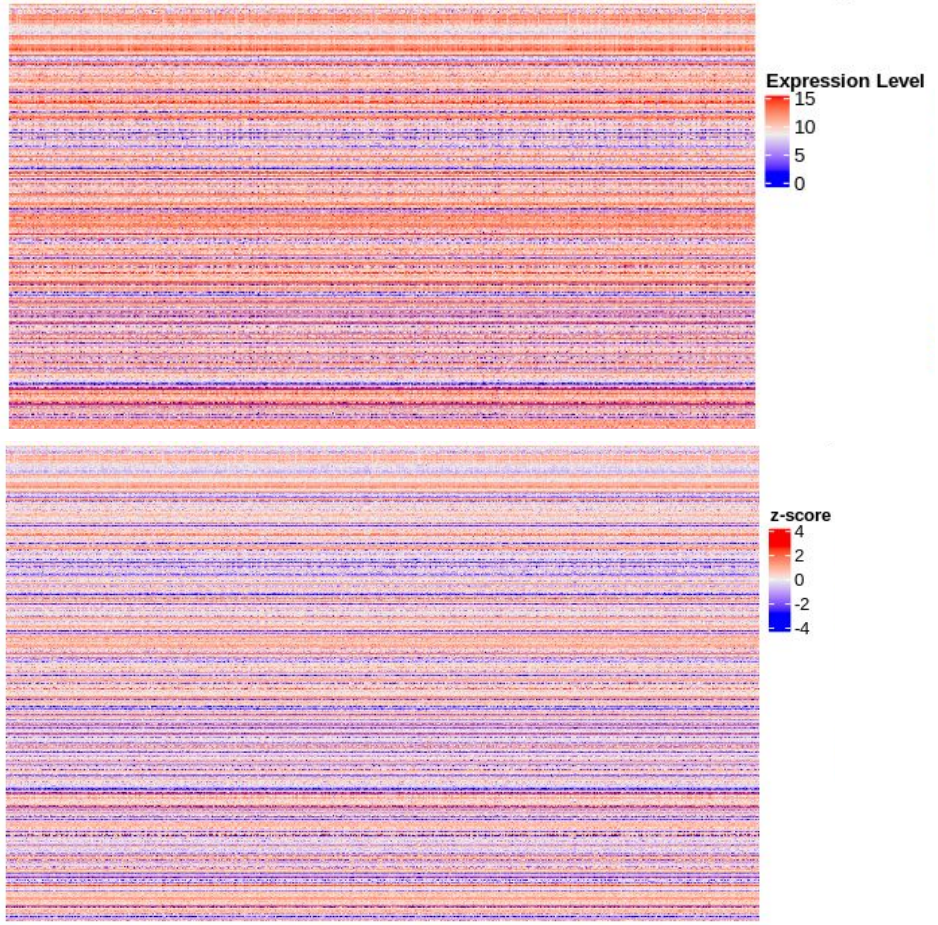
High-level library pheatmap, ComplexHeatmap (BioConductor)

for annotations and basic clustering : hierarchical, k-means...

R: > Heatmap(...,cluster_columns = TRUE)
> columnAnnotation()



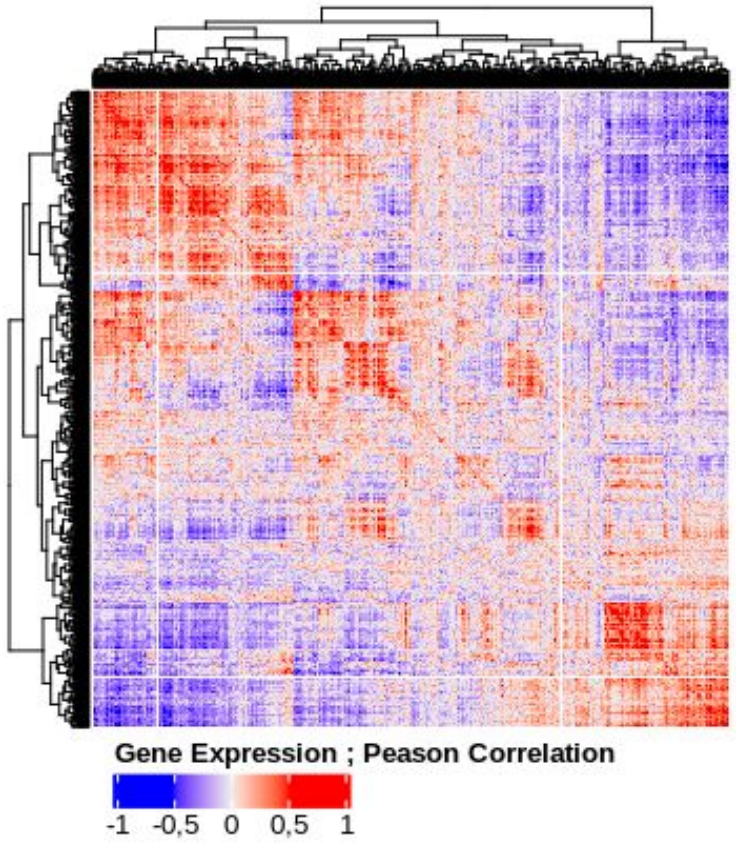
for z-score (columns) in R: scale()



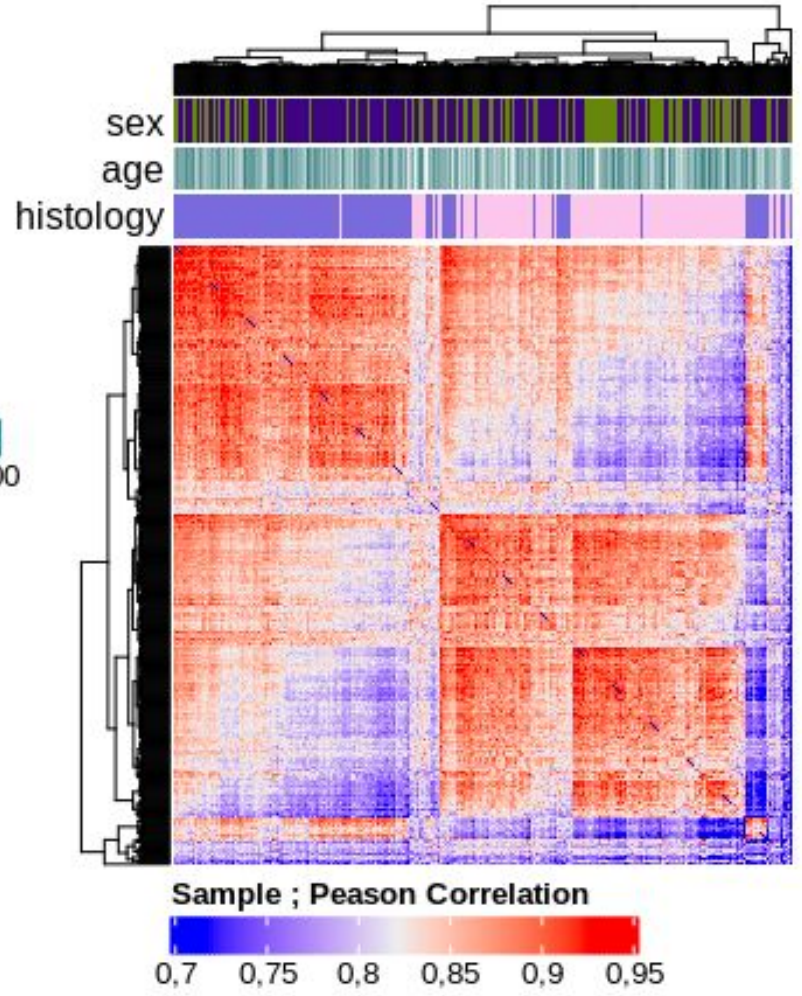
Data Visualization for Genomics Data : Co-expression and Correlation

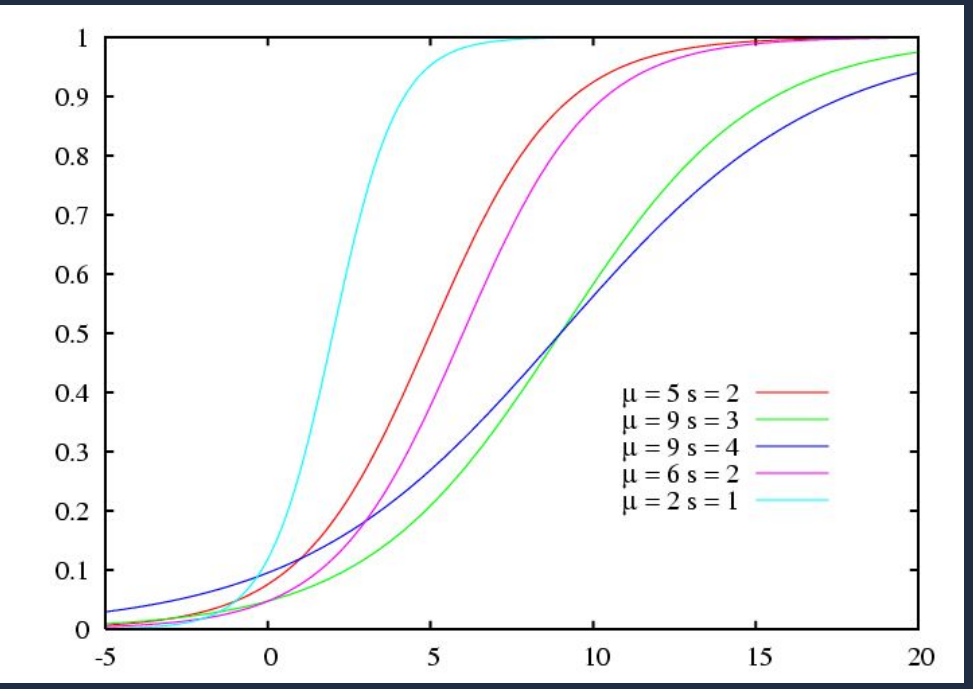
Pearson correlation and distance Matrix in R:

```
>df=as.matrix(as.dist(cor(df, method="pearson"))
>row.names(df)=c(); colnames(df)=c() ;
>Heatmap(df,cluster_rows=TRUE,cluster_columns=TRUE)
```



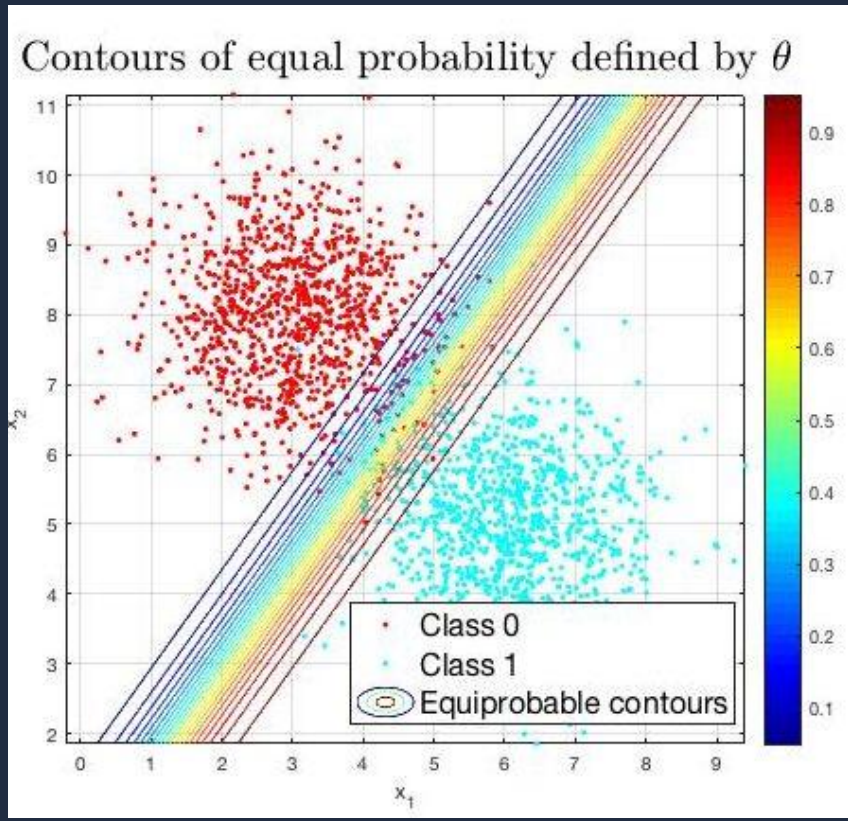
basic clustering of samples :





Logistic Regression

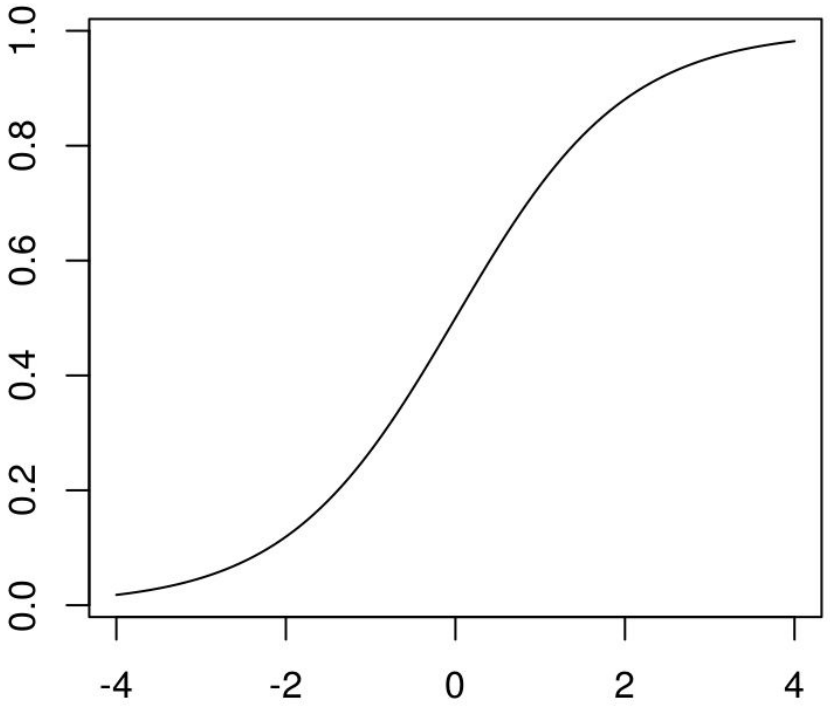
What is logistic regression ?



Logistic regression 1/3

Logistic regression - Linear Model

sigmoid function :



Goal : to attribute the class (1) for an observation x

$= \boldsymbol{\pi}(x) =$

For a vector of observations) we can use the vector formula similar to linear regression :

=)

R: `glm(y~x1+x2,family=binomial)`

Logistic regression 2/3

Logistic regression - Model building. Example with a series of ***nested models*** M

M(0) $\pi(x)=$)

M(1) $\pi(x)=$

M(2) $\pi(x)=$)

...

M(*p*) $\pi(x)=$)

adding variables : "forward"

removing vatiables : "backward"

Critere AIC : $2k$ -2 $ln($)

AIC in R : `extractAIC(m)`

in R : `logLik(m)`

R: function `step` calculate best model based on AIC

```
m_lo = glm(y~1,d,family=binomial(logit))
m_up = glm(y~.,d,family=binomial(logit))

m = step(m_lo,
    dir="forward",   scope=list(upper=m_up,
        lower=m_lo))
```

Logistic regression 3/3

Penalized Regression : + λ [/2 +]

Lasso: ; Ridge: ; ElasticNet :

Gaussian:

Binomial:

R: package `glmnet` (Hastie)

`m=cv.glmnet(x, y, family = "binomial")` : cross-validation for models with λ

`predict(m,xnew, type="response")` : fitted probabilities for all models λ

`predict(m,xnew, type="class")` : predicted classes for all models λ

`predict(m,xnew, type="nonzero", s)` : list of selected variables λ=s

`predict(m,xnew, type="coefficient", s)` : coefficients at λ=s

Advanced ML classification algorithms

random forest
xgboost
svm

Advanced ML - Decision trees

Supervised learning, mostly for labeled data

Nodes are *basic rules* on 1 variable :

defines *splits* : boundaries *in only 1 dimension*

Provides Non-linear boundaries

Training : Recursive Binary Splitting

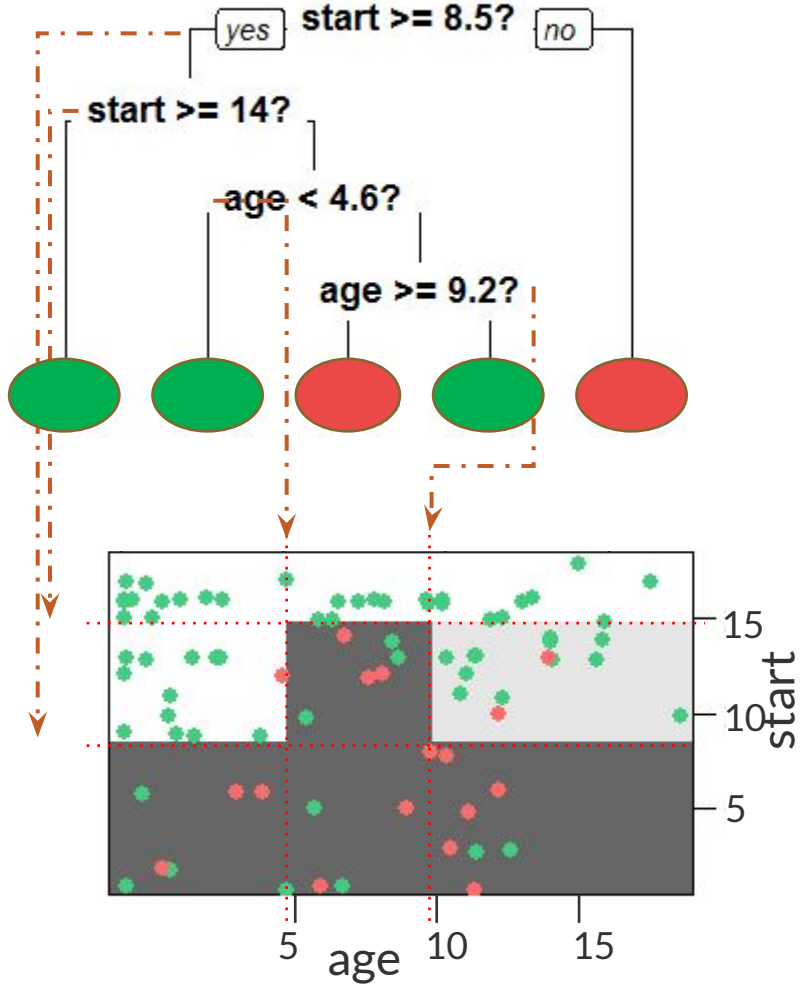
Number of rules : Depth

Depth too small: poor fitting

Too many rules : overfitting

Pruning: remove sub tree

balance complexity vs. fit



Advanced ML : Boosting

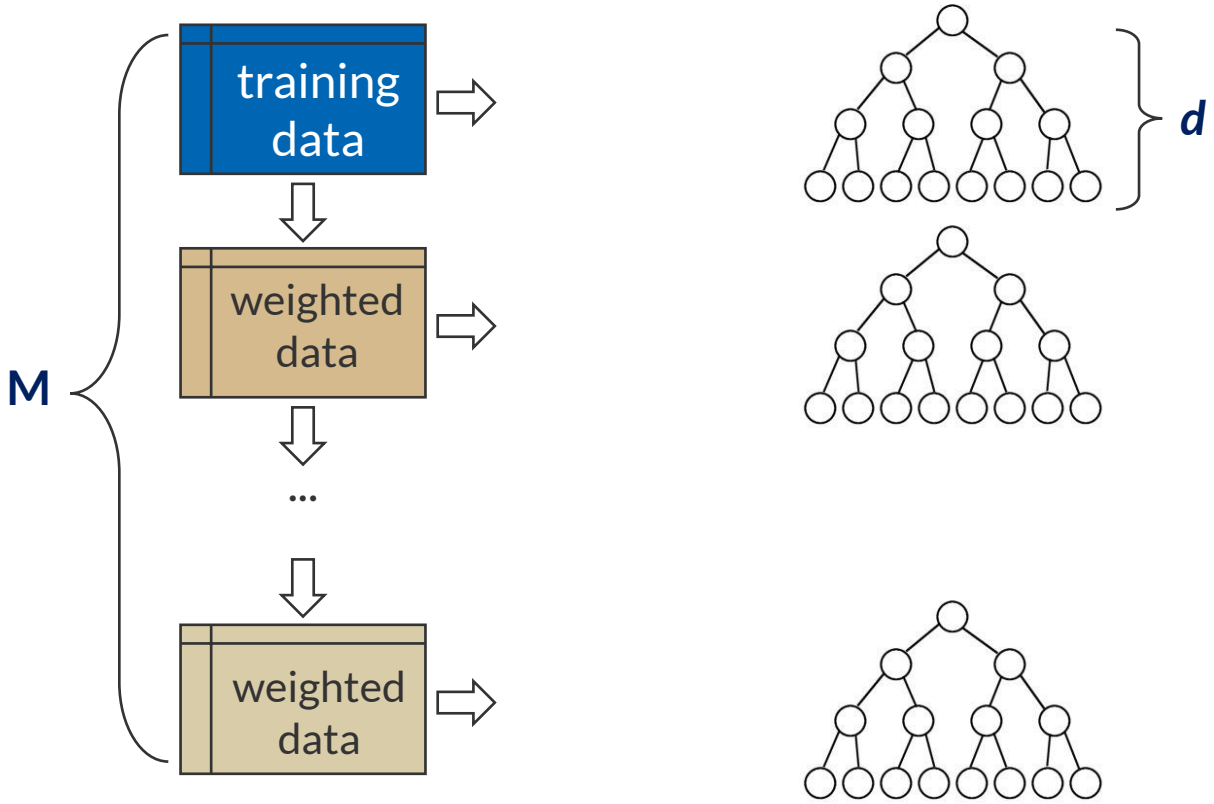
Package xgboost
example with algorithm ***AdaBoost.M1***:

- Define **M** trees of depth ***d***
- Misclassified samples at step ***m*** provide **weights** for data at step ***m+1***

;

- **Error rate** at step m provides coefficient for final classification :

= sign()



R: `xgboost(data,label,max.depth=d,nrounds=M,objective="binary:logistic")`

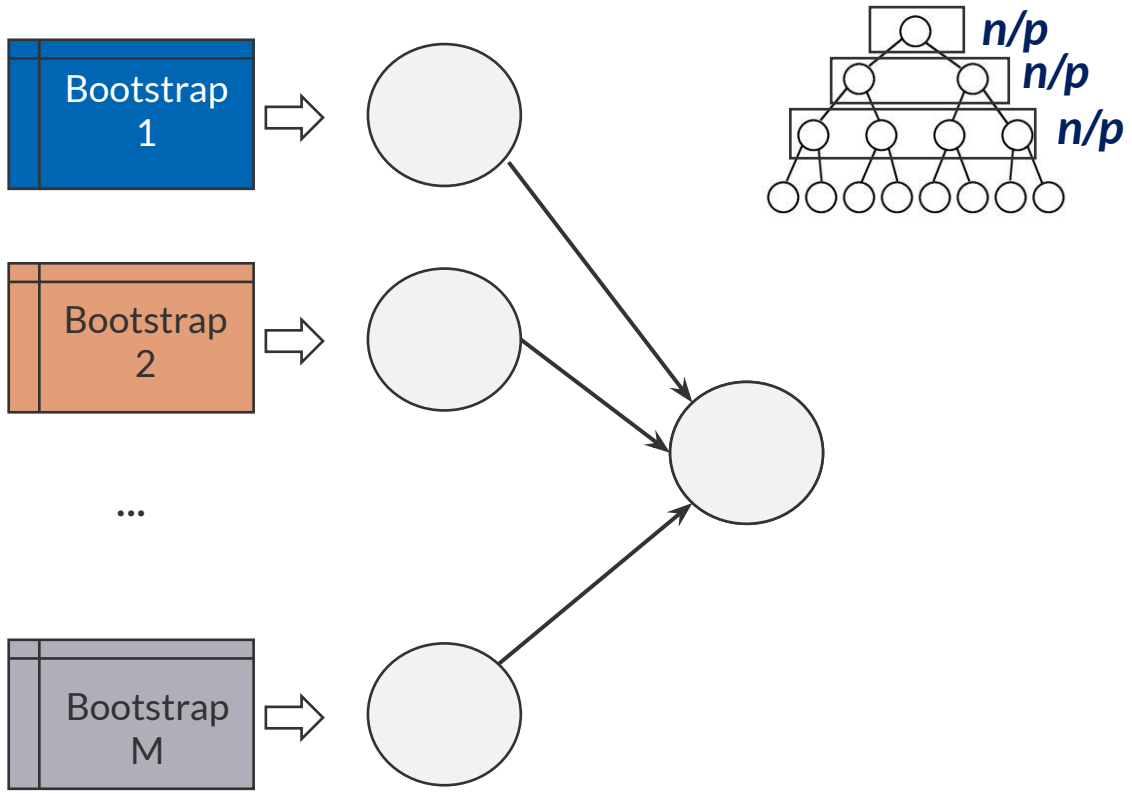
Advanced ML : RandomForest

Package RandomForest
extension of ***Bootstrap AGGregatING***

- Define **M** bootstrap datasets
- Train **M** trees of depth ***d*** with
  - **random trees : sample *n features*** (variables) before each split

- Final classifier averages predictions :

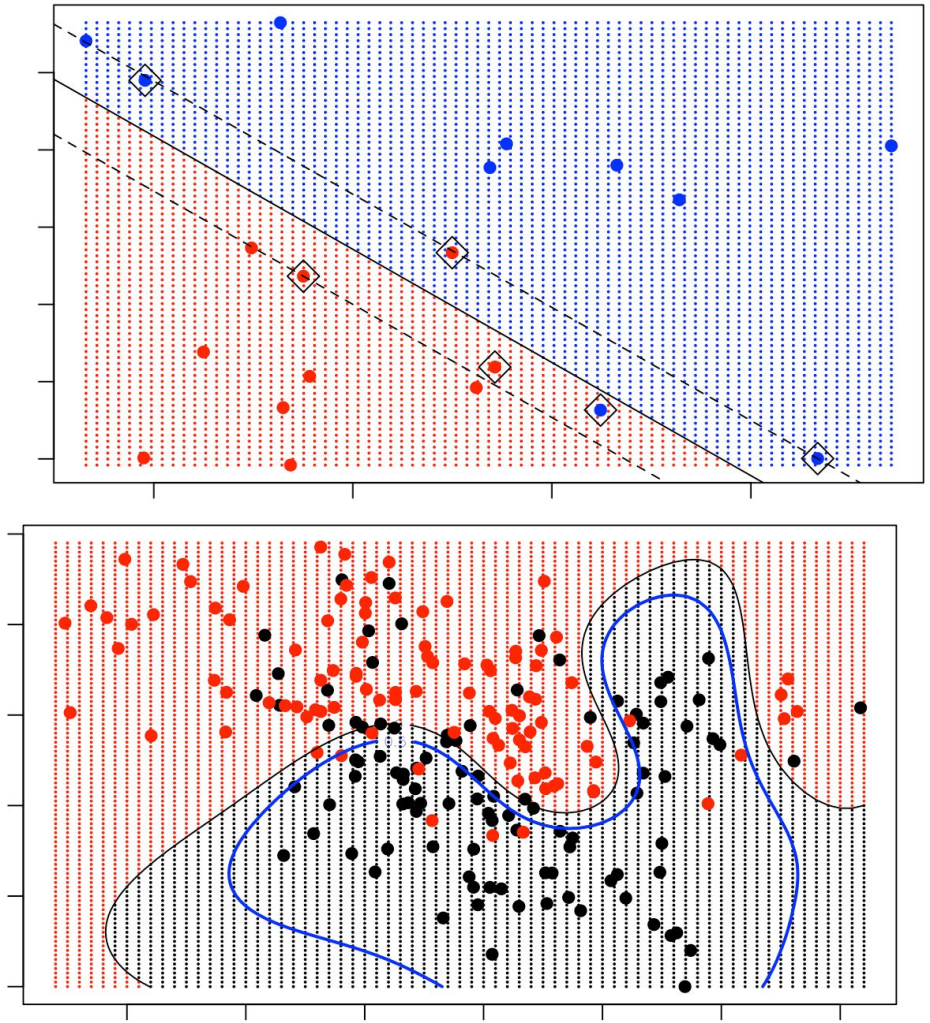
)



R: `randomForest(type=classification,x=data,y=labels,ntree=`**`M`**`,mtry=`**`n`**`)`
`M` : large enough (500-2000) ; **`n`** : regression : /3 , classification :

Advanced ML - SVMs

- Supervised method ; mostly for labeled data
- linear separation
  hyperplan :
- defines margins (dotted line)
  - **support vectors** are samples that are **within** the margin
  - optimal boundary will **maximize margins** (minimizing) **using support vectors**
- Transforms data for **non-linear boundaries**:
  - polynomial kernels :
  - radial kernels :



Advanced ML - Choosing Models

R:
`table(factor(training) , factor(prediction))` : Create confusion table

package `caret`
`createFolds(data$labels, k )` : create cross-validation k -folds

`train()` : high level creation and evaluation of many many many models

`confusionMatrix(factor(training), factor(prediction)` : Stats for confusion table

`pheatmap(data, annotation_row , show_rownames...)` : heatmap